

Can biological phenomena be understood by humans?

Despite the increasingly successful collaboration between physics and biology, there are contexts in which their styles and philosophies can diverge. Nowhere more so than in 'understanding'.

“It is nice to know that the computer understands the problem. But I would like to understand it, too.” So said the theoretical physicist Eugene Wigner when shown the results of a large quantum mechanics calculation. Now, if a theoretical physicist of a younger generation is to be believed, today's biologists are in danger of delegating to their computers the job of understanding biology. Physics Nobel laureate Robert Laughlin — who was able to understand the fractional quantum Hall effect without the aid of computers — was invited to a conference in San Diego earlier this month on quantitative challenges in the post-genome-sequence era. His job was to stir things up, and he duly did so. His remarks sparked a spirited defence of the post-genome-sequence biological agenda, but also prompted some useful contemplation of just what understanding means — or may come to mean — in biology.

To understand Laughlin's complaint, take the field of structural genomics. Here the agenda is to work out the structures of all the proteins encoded by an organism's genes, on the premise that a protein's structure can be a useful guide to its function. The problem is that protein structures are hard to obtain: the number of experimentally determined structures will never catch up with the number of known sequences. But there turns out to be great redundancy in the structures adopted by protein domains: only a limited number of structural arrangements (or 'fold' types) are found in nature. So the plan is to obtain the structures of all the fold types and then to model proteins of unknown structure into these folds on the basis of similarity of sequence, chemical plausibility and so on. In this way, it should be possible to model most globular protein structures with useful accuracy by the time the Human Genome Project has been completed in 2003.

Dangers of the black box

So far, so good, and this seems an eminently sensible way to identify new candidate drug targets, or to generate hypotheses about protein function that can then be tested by experiment. But if structural biologists then have a computational black box into which they can feed an amino-acid sequence, and get a protein structure out of the other end, can they say that they understand protein folding?

Laughlin would say no. He looks at a protein and sees an “emergent phenomenon”, an entity whose properties cannot be derived from the properties of its parts. So, just as the behaviour of a convecting fluid cannot be predicted from analysing its component molecules, so the function of a protein does not reside in the properties of its amino acids. Moreover, he argues that the higher-order organization (the protein's structure and function) is insensitive to the nature of the parts — the details don't matter. In one respect this is certainly true of proteins, in that totally different sequences of amino acids can yield the same fold structure, and thus the same function. But proteins can also be exquisitely sensitive to some subset of the details: change one amino acid (if it's the right one), and the protein function can be destroyed. The challenge is to find out which details count, and why.

Laughlin worries that in the post-genome era, biologists will be too busy accumulating facts and modelling them to seek simplification and underlying principles. The fact-gathering tendency is apparent not just in structural genomics, but also in functional genomics (where the aim is to identify the role of each gene in the genome) and proteomics (with a similar aim for each protein in the cell or organism). Evidently, there are enough facts to keep biologists busy gathering them for decades, so when will they have time to think?

In defence of modelling

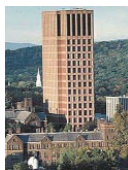
One answer, conveyed forcefully at the conference by Klaus Schulten (for protein-structure studies) and David Botstein (for genomics), is that modelling aids thinking, by helping to generate hypotheses. Where a theoretical physicist might gain intuition from pencil and paper, a structural biologist may need molecular dynamics simulations. (And, to be fair, physicists who work on complex systems are themselves no strangers to computer modelling.) The microarrays that are becoming the workhorses of genomics can produce a million data points in a single study; even a physicist, argued Botstein, would find it hard to make sense of these data as a table of numbers. So it makes sense to use the computer to put the data in some kind of order before trying to think about them. The ordering may be driven by a hypothesis — for example, one might look for evidence of a periodic cell cycle in the transcription of genes — but in the absence of a hypothesis, the data themselves may suggest a pattern.

Although one structural biologist went so far as to proclaim that “we have to free ourselves from the hypothesis-driven approach”, even he admitted that it would be nice to find some underlying simplicity. But he and others expressed the view that a search for unifying principles would be premature — that it will first be necessary for biology to go through a stage of fact accumulation and pattern recognition.

But what if it's not just a 'stage'? Is it possible that the parts will be enumerated and the functions found, and still there will be no simplification? Fortunately, enlightenment can come in different forms — not just in the elegant simplicity of a physicist's theory, but also in the more utilitarian guise of an engineer's analysis. As Hartwell *et al.* have argued in the recent supplement *Impacts of Foreseeable Science* (*Nature* **402**, suppl. C47–C52; 1999), molecular and cell biology may have more in common with engineering and computer science than with the basic sciences; for example, the kind of modelling needed to understand the complex intracellular networks that underlie most biological functions comes straight from engineering control theory.

As shown by two papers in last week's issue (see *Nature* **403**, 335–338, 339–342; 2000), it is becoming possible not just to analyse naturally occurring networks in this spirit, but also to design and build biological networks to implement desired functions. That, surely, is a kind of understanding worth having, and one that theoretical physicists can recognize as progress of a sort.





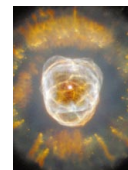
US universities
Yale plans to spend \$500m on boosting its image in science
p349



OncoMouse
NIH researchers get free use of DuPont's medical mouse
p350



Dolly patent
Roslin Institute awarded patents for cloning technology
p351



Eskimo nebula
Refurbished Hubble catches death throes of a planetary nebula
p355

All parties keen to press on with Europe-based science website

Paris

Plans to launch a Europe-based global website for the scientific literature, E-Biosci, were unanimously endorsed by research organizations, commercial publishers and the European Commission in Heidelberg last week.

There was less agreement on what form the proposed site should take, however. There was a strong feeling that it should cooperate rather than compete with the similar US initiative, PubMed Central, that is being led by the National Institutes of Health and is due to be launched next week (<http://www.pubmedcentral.nih.gov>).

Major publishing houses at the meeting included Elsevier Science, Springer, Blackwell and the Nature Publishing Group. "The presence of the publishers was critical," says Tony Mayer, an official at the European Science Foundation in Strasbourg.

The meeting also confirmed the differing US and European approaches to the goal of a single full-text searchable gateway for the scientific literature. PubMed Central (PMC) has made 'barrier-free access' its *raison d'être* and has invited commercial publishers and learned societies to deposit papers on its site free of charge — which one observer likens to "inviting turkeys to a Christmas dinner".

Several journals, including the *Proceedings of the National Academy of Sciences*, have signed up, and more are likely to do so (see over). But most commercial publishers have shunned the initiative. In contrast, the organizers of E-Biosci seem ready to compromise on barrier-free access, at least in the immediate future.

One proposal floated at last week's meeting by Stefan von Holtzbrinck, managing director of the Nature Publishing Group, was that publishers might cooperate with E-Biosci to establish a single full-text searchable site (see *Nature* **402**, 115; 1999).

The emerging picture is that E-Biosci would hold abstracts covering a range of disciplines, linked to the full text of the articles. Brigitte Volk-Zeher, an official at Germany's Deutsche Forschungsgemeinschaft, wants a site "for all abstracts in science that would be much more complete than PubMed".



von Holtzbrinck: "A global venture."

Such a site would be a valuable archive, says Frank Gannon, executive director of the European Molecular Biology Organization (EMBO) and the driving force behind E-Biosci. "But whether it would be free would be a 'per publisher' decision," he adds.

This scheme would fall short of the ideal of a "single, searchable, current, complete and free site," says Gannon, but it would let E-Biosci quickly acquire a critical mass of current papers and allow full-text searching across a swathe of journals. Ways to achieve freer access could be negotiated with time.

"We are talking about a portal for all life sciences, a global repository headquartered

in Europe," says Gannon. "The reality is: will anything happen if you do not reach a compromise with the publishers?"

"One way to potentially break the stalemate between the publishers and these initiatives is to take the positive points and begin in small steps," says von Holtzbrinck. "The 'E' in E-Biosci stands for electronic, not Europe," he adds. "It is a global venture."

Some feel that Gannon has gone too far in his efforts to meet publishers' needs. The cost of a freely accessible server "would only be a fraction of what is now disappearing into the pockets of publishers' shareholders," says Raf Dekeyser, chief librarian at Leuven University in Belgium. He criticizes "the influence of the publishers and some scientific societies who are afraid to lose their source of income".

"In my opinion, the big players like Elsevier will never voluntarily commit suicide by

Quality, not quantity, for UK PhDs?

London

Annual stipends for British research students should increase by more than a third — even if it means fewer studentships — according to a report published last week.

The report was produced for the UK Life Sciences Committee, an umbrella organization representing a range of biological and biomedical societies. It calls for better quality and support of research studentships, rather than maximizing the number of students at the lowest price.

"Filling the places is not the same as getting the best students into science," says Brian Follett, vice-chancellor of the University of Warwick and chair of the group that wrote the report. The report says that all PhD students funded by the UK research councils should receive a tax-free stipend of at least £9,000 (\$14,900). The average is currently about £6,620.

Follett says he hopes the government will provide fresh leadership and thinking on PhD funding and recruitment — an area

that has been unchanged for at least 40 years. The timing of the report is appropriate, he says, given that the government is in the middle of a public-spending review.

Policy changes mean that new graduates carry debts of around £10,000. The PhD system must change to reflect this, says Follett, because the average research student would not clear such debts until 15 years after graduation. This would deter people from poorer backgrounds from entering research.

Follett acknowledges that not all academics will be pleased to see a reduction in the numbers of PhD students. But, he says, "we need a culture change, away from a policy of maximizing numbers towards a much greater emphasis on the student, their training and their future."

He says that he is optimistic that government officials will take the report on board. It is now with the director-general of the research councils and the chief executives.

Natasha Loder

giving their content away for free," says Dekeyser. "We have to force them to come down to reasonable prices by first creating an alternative electronic distribution channel, with the help of the refereeing scientific societies, in order to guarantee quality."

E-Biosci has been weakened by publisher involvement, he says. "In spite of the enthusiasm, the final E-Biosci proposal is very weak." Others point out, however, that initiatives such as PMC and E-Biosci are giving publishers sleepless nights, and the fact that Elsevier — notorious for its 40 per cent plus profit margins — attended last week's meeting is significant.

The days of fat profit margins are over, says one publisher, arguing that E-Biosci allows publishers to evolve without driving them to extinction. He predicts that publishers will settle for 10 per cent margins, while seeking new routes of income, other than primary publishing, through on-line opportunities.

The European Bioinformatics Institute (EBI) presented a bid to host E-Biosci. But while participants seemed enthusiastic that EBI should play a key role, many felt that a decentralized network was preferable, with EBI as the central node in a European network and other nodes in the member states.

One official from the European Commission says that the meeting's main legacy may be a pan-European information-technology infrastructure, linking existing electronic libraries such as DIMDI in Germany and France's INIST.

DIMDI already loads databases from the US National Library of Medicine as well as European ones, points out Dieter Kaiser, heads of its System Development Group.

All the representatives of member states' research councils contacted by *Nature* said that they strongly backed E-Biosci, but could not make financial commitments until the proposals had been fully discussed at home.

The inherent delays in reaching agreement among European member states could be the biggest obstacle to making speedy progress, warn several of the participants, who are concerned that the project risks getting bogged down in endless discussions.

Although little money is on the table, E-Biosci's estimated annual budget of 3 million euros (US\$3 million) is small, says Marja Makarow, who represented the Academy of Finland and the country's Medical Research Council at the meeting. E-Biosci could allow national funding agencies to make economies in library costs, Makarow adds.

A recurrent theme was whether public money might be better spent on transferring resources from conventional libraries to a global archive. Librarians and universities urgently need to be brought into discussions of E-Biosci, say many participants.

The meeting agreed to submit a proposal to the European Commission for funding,



EBI: candidate for a central node.

but felt that money was needed immediately to launch the initiative in the coming months. It agreed to set up a task force to resolve technical requirements for E-Biosci, while the publishers will discuss their joint initiative further.

The European Molecular Biology Council, which represents member states, will seek to coordinate funding for E-Biosci through its member states, with input from the European Science Foundation, which covers a wider range of disciplines. But Carthage Smith, head of international relations at Britain's Medical Research Council, says it is "far more important for Europe to get this right than to do it quickly".

The mood of the meeting was that a European archive was essential to avoid a US monopoly on a global scientific archive. But it was also unanimous that E-Biosci should collaborate, rather than compete, with PubMed Central. "We need coordination with the US, but coordination of equals," says Glauco Tocchini, secretary general of the European Molecular Biology Council.

"Anything that Europe does must be coordinated with the United States; if we end up in a competitive situation over the scientific literature then it is to no-one's advantage," says Smith. "Europe can go in

with a different starting position but all the content must be shared. It must be a global resource."

David Lipman, director of the US National Council for Biotechnology Information, which operates the PubMed and GenBank databases, says he is "hopeful that we will set up a smooth collaboration" and adds: "in any case, we need advice and input from European scientists and administrators for PubMed Central".

PubMed Central will publish refereed material from existing journals or new organizations such as Current Science's BioMed Central; it will also have a server for unrefereed e-prints from academic institutions. Most of the participants at the meeting insisted on the need for peer-review, and the original E-Biosci proposal for including both peer-reviewed material and preprints was widely contested.

In particular, they rejected an EBI proposal that the archive should contain articles that had simply been checked for accuracy. Smith describes this concept as "dangerous, pseudo quasi review," adding: "We have enough trouble reviewing the reports for the best journals. Scientists are not going to review all this stuff."

"A repository for PhD theses within E-BioSci should be established," adds Henning Beck-Nielsen, chairman of the Danish Medical Research Council. "This will also promote a harmonization on the level of quality for PhD degrees." **Declan Butler**

Yes, no and maybe: where they stand on PubMed Central, the US biosciences online archive

<i>Proceedings of the National Academy of Sciences</i>	Content available four weeks after print publication, for an experimental period of one year; contingent on non-peer-reviewed material being kept off PMC. No third-party use.
<i>Molecular Biology of the Cell</i>	Full contents available next year on an experimental basis, two months after publication. Will also provide back issues.
<i>Canadian Medical Association Journal</i>	Agreed to join.
<i>Frontiers in Bioscience</i>	Agreed to join.
Current Science Group	Five journals available initially; also providing peer-review services for PMC (see <i>Nature</i> 402, 110; 1999)
<i>British Medical Journal</i>	Content available at time of publication.
<i>Journal of Cell Biology</i>	May post 1–2-year-old issues, contingent on peer-reviewed journals not being mixed with an unreviewed preprint server. Also interested in "independent non-profit and commercial solutions to the linking problem".
American Physiological Society	Will offer content free 12 months after publication; will not release before without a viable financial model.
American Society of Microbiology	Supports concept. Makes content freely available via HighWire after one year. Does not intend to participate in PubMed Central until the requirement for hosting on the NIH server is dropped.
<i>EMBO Journal</i>	Undecided. May join as part of a joint venture between PMC and E-Biosci (see above).
<i>Genes and Development</i>	Believes that "further study is needed" and many questions still need to be answered.
<i>Journal of Clinical Investigation</i>	Provides free access via HighWire. "A three-year 'experiment' with free access has been associated with a 10 per cent drop in overall subscriptions." Willing to consider full text access via links to HighWire; moving to PMC "contingent on the level and quality of service being at least as good as HighWire."
<i>Journal of Neuroscience</i>	In discussions; undecided.
<i>Nature</i>	Agreed that something like PMC could help enhance <i>Nature's</i> ability to serve the research community, but sees significant unanswered questions. In discussion with PMC and E-Biosci.
<i>Nucleic Acids Research</i>	Still negotiating, but expects to join PMC early this year.
Oxford University Press (OUP)	"In negotiation with the National Council for Biotechnology Information about the possibility of OUP submitting some journals to PubMed Central." Wants "same kind of usage data that we currently get from the HighWire system".
<i>Science</i>	"So long as the NIH insists that it must have our meta content for both searching and dissemination, we will stand by our current decision not to participate."

Clinton proposes \$2.8 billion increase in science funding

San Diego

President Bill Clinton has proposed a \$675 million boost to the US National Science Foundation's (NSF's) annual research budget — the largest increase ever. The money forms part of a \$2.8 billion increase in the overall US research budget.

Clinton announced his plan for increases in his administration's 'Twenty-first Century Research Fund' in a speech at the California Institute of Technology in Pasadena last week. Some US scientists see this as evidence that the White House has heeded people such as NSF director Rita Colwell, who have sought more investment in basic scientific research.

The fund includes an additional \$1 billion for the National Institutes of Health (NIH), whose budget stands at nearly \$18 billion. Scientists were encouraged by the proposed increase, but cautioned that it is less than they hope the agency will receive. It is not on track to double the research budget, as some science leaders want.

The Association of American Medical Colleges and the Federation of Societies for Experimental Biology (FASEB) are both seeking a \$2.7 billion increase in the NIH budget. This is designed to double the NIH budget within a five-year period which began two years ago.

"This is a fine initial proposal," says David Kaufman, a pathology professor at the University of North Carolina and FASEB president. Last year Clinton proposed only a \$300 million increase in the NIH budget.

Clinton's proposal includes a \$497 million National Nanotechnology Initiative (see *Nature* **400**, 95; 1999), along with a \$600 million increase for research into information technology.

White House officials say the nanotechnology project would double federal spending on this research. The president's proposal calls for participation from the NSF, NIH, the Departments of Defense, Commerce and Energy, and the space agency NASA.

Rita Colwell says that the record dollar increase for the NSF "will give us the capacity to make strong across-the-board investments in science and engineering research and education". The foundation currently has an annual budget of \$3.9 billion.

Science leaders see the timing, focus and location of Clinton's initiatives for the fiscal year 2001 as significant. He made the announcement close to important centres in biotechnology and information technology, on the way to Democratic fund-raising events in a state that is crucial for this year's presidential election.



Driving force: Clinton characterizes science and technology as "the engine of economic growth".

Unusually, the research budget was announced separately from other budgetary plans. Clinton's speech highlighted the connection between research and the economic boom. "The first thing I want to underscore, in the clearest possible way, is that science and technology have become the engine of economic growth," he said.

"The president has indicated that funding university research is a top priority," says Mike Lubell, a physicist at the City College of New York, and spokesman for the American Physical Society. "That is a very strong statement for the future."

But although the NSF is to get twice as much as its previous largest increase, Lubell noted that, on a national scale, research funding still lags behind the heyday of the 1960s. Then, the federal research budget, as a percentage of gross national product, was twice what is now proposed. "One could argue that the prior investment paid off tremendously," says Lubell.

Republicans in the House of Representatives have questioned how Clinton's proposal will be funded, given his previous record. Two years ago, an increase in the research budget was predicated on an increase in tobacco tax that did not become law. And last year, Republicans criticized research budget increases for being funded by 'gimmicky' increases in taxes and fees.

James Sensenbrenner, the Republican chairman of the House Science Committee, said he was cautiously optimistic. But he added that he couldn't determine whether the proposal means a higher priority for science until details of the president's entire budget are revealed.

"I am hopeful that these encouraging words reflect the administration's actual priorities, and are not merely promises within the context of an across-the-board government spending spree," comments Sensenbrenner.

Rex Dalton

Yale hopes a \$500 million boost will raise research profile

Washington

Yale University is to spend \$500 million over the next six to eight years on constructing and renovating its science and engineering facilities, in an effort to boost its reputation in these disciplines.

Private colleges across the United States are spending hundreds of millions of dollars on science and engineering buildings as they compete for students and research funding. "These schools are all vying for excellence," says Peter Smith, director of public affairs at the Association of American Universities. "In science and engineering, excellence requires facilities."

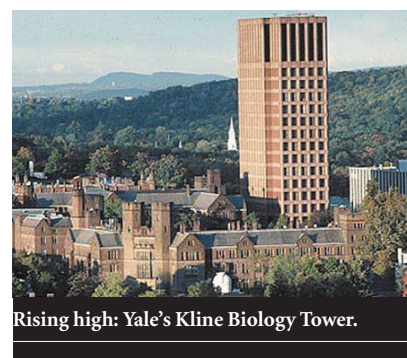
That may be especially true for Yale. Its president, Richard Levin, admits its reputation — excellent in the humanities — is weaker in science and engineering.

Yale already has some of the strongest science departments in the United States. The National Research Council ranked Yale's research doctorate programmes thirteenth in physics, twelfth in chemistry and tenth in cell biology among participating institutions in its 1995 report *Research-Doctorate Programs in the United States*. But the university did less well in engineering. The report ranks Yale's chemical engineering programme at position 32.5 and its electrical engineering at 30.5.

Whether the reputation is fair or not, the building campaign is an attempt to correct that perception, says Levin. New and improved buildings can attract better faculty, he adds.

The building programme may also represent a scheme to address years of deferred maintenance. The university began renovating campus buildings a decade ago, but has left its science buildings largely untouched. Some \$300 million will go towards renovating existing facilities; the remaining \$200 million will pay for five new buildings.

By spending a large sum of money



Rising high: Yale's Kline Biology Tower.

AP

STEVE DUNWELL

now, rather than smaller amounts on maintenance and renovation over the past decade, Yale is playing "a bit of catch-up" says Jeremiah Ostriker, provost of Princeton University. "I think they realized it was a necessity."

He adds that Yale will be able to compete in most areas of science. But catching up with some universities in engineering research may be more difficult. According to figures from the National Science Foundation, Johns Hopkins University spent \$213 million on engineering research and development in the financial year 1996; Yale spent \$6.6 million.

Princeton is investing \$200 million in construction, with about 60 per cent of it going towards new buildings, and a further \$100 million on faculty endowments and instrumentation, including \$55 million to set up a new genomics centre by 2002. Harvard University is to spend \$200 million on science and engineering building and renovation over the next five years.

David Baltimore, president of the California Institute of Technology, says there is "a national trend" of universities constructing science buildings. The institute will spend \$100 million on a new biology building starting in the spring; the building will cost \$40 million, with the rest going on instrumentation, renovation and faculty endowments.

The US economy seems to be fuelling some of the building boom, with computer-industry entrepreneurs playing a major role. The Massachusetts Institute of Technology, for example, has begun fund-raising for a computer-science and artificial-intelligence complex. It has already received pledges worth \$45 million from two donors, including \$25 million from Bill Gates, the chairman of Microsoft. **Paul Smaglik**

NIH cancer researchers to get free access to 'OncoMouse'

Washington

After four years of negotiations between the US National Institutes of Health (NIH) and the pharmaceuticals company DuPont, NIH-funded scientists now have free use of the 'OncoMouse', a transgenic animal technology used to create mice that develop tumours.

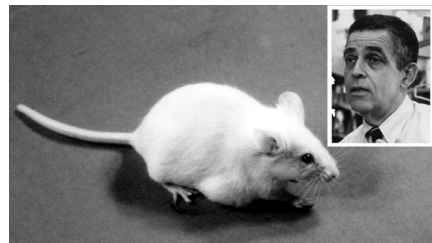
OncoMouse is the second research tool that the company has made public in the past two years. The deals are a victory for scientists who have argued that broadly applicable techniques should be available without strings to not-for-profit researchers.

As a condition for the use of both tools, DuPont originally required licences, demanded 'reach-through' rights on any inventions resulting from their use, and placed limits on breeding and redistributing animals that were altered with them.

But the NIH and DuPont set the precedent for the OncoMouse agreement with a 1998 agreement eliminating such terms for not-for-profit research using Cre-lox, a technology that allows researchers to remove genes from specific cells and tissues (see *Nature* 394, 819; 1998).

The OncoMouse and Cre-lox agreements are very similar. Both also adhere to the NIH's principles for sharing research tools (see *Nature* 403, 10; 2000). Together they are "signposts for how similar problems might be solved", says Harold Varmus, president of the Memorial Sloan-Kettering Cancer Center in New York. Varmus helped negotiate with DuPont while he was NIH director, a position he held until the end of last year.

Varmus stresses that both agreements distinguish between products, such as restriction enzymes, that are "consumable", and



Public mouse: the NIH–DuPont deal has lifted a cloud of fear. Inset: Leder, OncoMouse creator.

techniques to make "things that replicate". Before the OncoMouse agreement, some cancer researchers operated under that distinction anyway, using the technique to develop cancer mouse models even though they had no licences or signed agreements with DuPont—a situation Varmus describes as "uncomfortable".

Those researchers feared that they were inadvertently infringing the patent, says Maria Freire, director of NIH's Office of Technology Transfer. She believes this agreement will lift that cloud of fear.

Both the deals differentiate between commercial and not-for-profit research. Commercially funded scientists must pay DuPont if they want to use either technique. But non-profit researchers, who can freely exchange animals altered with either technology, must alert DuPont if they distribute such animals to a commercial company.

The wider availability of the technique will allow more and better mouse models of cancer, says Phil Leder, a Harvard University researcher who developed the technology. "Cancer-prone mice are used in many research settings," he says. "This can only mean they can be more accessible." **Paul Smaglik**

Canadian biomedical collaboration keeps on growing

Montreal

The Medical Research Council of Canada (MRC) and the country's research-based pharmaceutical companies have announced significant increases for the second phase of a joint research programme, which has quadrupled in size since its introduction in 1993.

The collaboration's budget has risen from Can\$10 million (US\$7 million) at most each year at its beginning to more than Can\$40 million today.

Phase I has generated Can\$237 million for health research. The second five-year phase will introduce several new and

improved programmes, and the proportion of funding by the partners will change.

For training and salary awards, the MRC contributed Can\$1 for every Can\$4 from industry in the first phase; each will contribute half the cost in the second phase. In funding operating grants, the MRC also contributed Can\$1 for every Can\$4 from industry, but its second-phase contribution will be Can\$2 for every Can\$1 from industry.

New operating fellowship awards of up to Can\$49,000 a year for three years will be available for scientists in clinical investigation and interdisciplinary research, plus up to Can\$30,000 in research allowances in

the first year. University–industry research chairs will also be created, at a cost of up to Can\$140,000 a year plus operating funds.

Funding ratios for clinical trials have been left at Can\$1 from MRC to Can\$4 from industry. But provision will be made for 'add-on studies' to allow additional research, with infrastructure costs being absorbed by the larger study.

Marc Lepage, a spokesman for the MRC, says that, although it is impossible to estimate the total amount that the second phase of the programme will generate, the MRC would be "very disappointed" if they "only made \$237 million". **David Spurgeon**

Ministries cooperate to plan Japanese genome centre

Tokyo

Japan's struggling genome research sector will be boosted by the opening of a new centre, the result of cooperation between several ministries. The centre's director will coordinate the country's research on the application of single-nucleotide polymorphisms (SNPs).

The centre, for which the Science and Technology Agency (STA) has set aside ¥1.9 billion (\$18 million), will be located in Yokohama at Riken's Genomic Sciences Centre. In the past, Tokyo University's Institute of Medical Sciences and Tokushima University accounted for most SNP research in Japan.

Under Prime Minister Keizo Obuchi's Millennium Project, more than 20 universities and several foundations, consortia, and public and private research institutions will work together in an initiative sponsored jointly by STA, the Ministry of Education, Sports and Culture (Monbusho), the Ministry of Health and Welfare (MHW) and the Ministry of International Trade and Industry (MITI).

The project will draw together several strands of research, from the sequencing of 30,000 full-length complementary DNAs and the identification of 150,000 SNPs, to SNP frequency analysis, construction of reference databases, diagnostic screens, and pharmaceutical and medical applications.

Kumao Toyoshima, of Sumitomo Hospital in Osaka, is tipped as the probable new director. He will be expected to coordinate research that leads from SNP databases to practical applications in five key areas: cancer; metabolic disorders such as diabetes; circulatory disorders such as high blood pres-

sure; immunological disorders; and dementia. This will mean working closely with Tokyo Medical Centre and the MHW's National Cancer Centre.

The centre is well funded, but some people are concerned that the involvement of several ministries and institutions may lead to inefficiency. One senior researcher criticizes what he calls a "budget first, strategy later" approach in which money is first allocated to the ministries, which decide how to distribute funds among their own institutions.

Critics say that a better alternative would be to determine the research needs of the various institutions in an interministerial committee and then allot money accordingly. The top-down strategy also leads to some unnatural divisions in research structure, such as between the identification of SNPs and the search for disease-related genes.

Funding in the new plan is focused on identifying SNPs, rather than medical applications. One senior researcher speculates that structuring the project as five divisions probably stems from demands by the five ministries to appoint a division leader each.

The success of the project will depend on researchers' ability to overcome bureaucracy and integrate their research. But many are hopeful that they can do this, and enthusiastic about the government's recent push into SNPs research (see *Nature* 399, 295; 1999).

Masaaki Terada of the National Cancer Institute believes that researchers will coordinate their efforts to avoid doubling up on research projects. "They are competitive, but they do not want to waste their time doing the same thing as someone else."

David Cyran

Issue of patents on 'Dolly' technology stirs controversy

Washington

The first patents on the technology used to clone Dolly the sheep were issued last week, three years after its inventors published news of Dolly's birth (see *Nature* 385, 810–813; 1997).

The two British patents, issued on 19 January to the Roslin Institute and the two government agencies that supported its work, cover a broad variety of applications of the cloning technology, but exclude human reproductive cloning. In all, the patents cover 58 claims to methods of producing cloned non-human animals, human cell lines and early human embryos, and to the animals, embryos and cell lines produced in this way.

Other patents have already been issued for cloning technology that differs from the Roslin method, but covering only animals. The Roslin patents cover the cloning and growth of a human embryo to the level of the blastocyst, an early embryonic stage.

Separately, the US Patent and Trademark Office has issued a notice of allowance for a patent application on Roslin's cloning technology, also called nuclear-transfer technology. A US patent on the technology is therefore likely to be issued within months.

The US patent is limited to techniques for cloning non-human mammals, but the inventors are expected to follow this up with an application encompassing human uses.

"This is the first time anyone has had a patent issue on a nuclear-transfer technology that would cover its use in human cells," says David Earp, vice-president of intellectual property at Geron Corporation of Menlo Park, California. He says this puts his company in a "dominant position".

Geron, which last year acquired Roslin Bio-Med, a company set up by the Roslin Institute to develop its cloning technology (see *Nature* 399, 92; 1999), holds an exclusive licence on the Roslin patents and underlying technology, with one exception — PPL Therapeutics Ltd holds a licence to use the technology to produce certain human proteins in the milk of ruminants and rabbits.

Geron hopes to combine the cloning and stem-cell technologies to develop therapies for a range of diseases from diabetes to Parkinson's disease.

The idea is to use nuclear-transfer technology to produce a genetically

Top physicist quits Japan for US

Tokyo

Shuji Nakamura, who pioneered the development of blue light-emitting diodes (LEDs), is leaving Japan to take up an academic position in the United States.

Nakamura worked on gallium nitride-based LEDs and laser diodes at Nichia Chemicals, a privately owned manufacturer of fluorescent materials. He is taking up a professorship at the University of California, Santa Barbara, starting in February.

Following the demonstration of a workable blue laser, now commercially available, Nakamura says there was little left for him to achieve at Nichia.

According to Nakamura, many industrial researchers are leaving Japan for better work-



Nakamura: leaving for better conditions.

ing conditions in the United States. "Japanese industrial research and development may be on its way to becoming obsolete."

Nakamura, who has not made up his mind about future research, says he did not receive any offers from Japanese companies or universities, but cannot imagine working for a Japanese university. "Because of numerous regulations, Japanese universities do not provide an environment where researchers can work freely."

Robert Triandl



Trouble ahead? Protesters have criticized the issuing of patents for cloning technology.

identical embryonic stem cell from a patient's own cell, then to cause the cell to differentiate into the kind needed for therapy — pancreatic cells in diabetes, for instance, or dopamine-producing neurons in Parkinson's disease. The potential therapeutic applications leave Earp "excited". "There are huge medical advances at our fingertips," he says.

The patents were criticized last week from various sides. "Organizations, including charities, may not have the money to pay to use this patent technology," Vivienne Nathanson, head of science and ethics at the British Medical Association, told BBC Television. "So this could inhibit medical research."

"It's going to break the health-care system," adds Jeremy Rifkin, who heads the Washington-based Foundation on Economic Trends. "If Geron and Roslin can get a lock on stem cells and [cloned] embryonic cells, they then have total control over future medical developments in this area of research."

Rifkin says he is distressed by the UK patent office's "chilling" and "unconscionable" decision to grant a patent on the earliest form of human life. He argues that patents should pertain only to inventions, not to naturally existing entities, such as humans. The decision, he says, is "an absolutely transparent violation of the patent statutes in virtually every country. This is something my attorneys are going to move on immediately with formal legal protests in the UK."

But Arthur Caplan, director of bioethics at the University of Pennsylvania, says Roslin has rightfully won the patents. "They've got the products and the uses as well as the technique." He adds: "It's time to shift the issue to a discussion of what responsible [patent] ownership is."

Geron's Earp says the company has "absolutely no intention" of inhibiting medical research. "We want to openly encourage as many people as possible to do basic research" using cloning and stem-cell technologies. **Meredith Wadman**

US survey reveals location of human tissue samples

Washington

A survey to be published next week will show the whereabouts of more than 300 million human tissue samples archived in the United States. The survey's authors hope it will form the basis for a comprehensive database of samples and increase their availability to the growing army of researchers who need them.

The Handbook of Human Tissue Resources, to be published by the Science and Technology Policy Institute at the Rand Corporation, attempts to provide comprehensive information on the hundreds of human tissue archives. These range from tens of millions of samples held by the Armed Forces Institute of Pathology to scattered collections in university departments across the United States.

"It is difficult for researchers to know what samples are out there, and what they need," explains Roger Aamodt, chief of the resources development branch at the National Cancer Institute (NCI). Aamodt's work compiling data on tissue samples from cancer patients helped form the basis for the broader Rand survey. "For the first time, the handbook puts this information in one public document."

Aamodt says there is "an incredible growth curve" in demand from researchers for human tissue samples, most of which were collected by pathologists without any intention that they would be used for research. But progress in genetics is turning these collections into powerful research tools. The Cooperative Human Tissue Network, which was established by the NCI in 1987 to obtain tissue samples for cancer researchers, distributed 50,000 samples last year to more than 500 investigators.

Aamodt suggests that the guide will be particularly useful for researchers at smaller institutions lacking specialist facilities to help their scientists find samples.

Rand started gathering information on US tissue sample archives for the National Bioethics Advisory Commission when it was preparing guidelines governing the circumstances in which the samples could be used in research. Many patients had not given informed consent to their samples being used in research, and there are concerns about the privacy of the genetic information they contain. The commission published its guidelines last year, and Rand has decided to publish the survey's results as a research tool.

Elisa Eiseman, the survey's chief author, says it should serve as a resource for basic and clinical research into many diseases. She adds that Rand is talking to the NCI and the International Society of Biological and Environmental Repositories about putting the survey's contents on the Internet. **Colin Macilwain**



India targets extra research funds

New Delhi

A \$500 million annual increase for the next five years in funding for Indian science and technology, promised by prime minister Atal Behari Vajpayee, will be used to boost basic research, modernize laboratories, and launch new technology missions, according to officials in the science ministry.

Valangiman Ramamurti, secretary for the Department of Science and Technology, says the news "has come at a time when funding for even routine basic research was steeply on the decline and several good proposals from universities had to be turned down".

Earlier this month, Vajpayee said that his government, led by the Bharatiya Janata

Party, would provide funding worth 1 per cent of the gross domestic product for research and development in the financial year beginning 1 April 2000, increasing to 2 per cent in 2004 (see *Nature* **403**, 126; 2000).

Ramamurti says that the heads of scientific agencies who met with finance minister Yashwant Sinha were given the impression that the increases would be reflected in the fiscal year 2000 budget.

According to Ramamurti, the additional resources would be spread across various scientific agencies under different ministries. At least 30 per cent would be used to boost basic research, with 20–30 per cent to go on "infrastructure creation". **K. S. Jayaraman**

Head of US watchdog faces uncertain future

Washington

The controversial head of the US government office that oversees the protection of human research subjects is likely to find himself out of a job in the coming months, when the office is dissolved and reconstituted in the office of Donna Shalala, the Secretary of Health and Human Services (HHS).

Gary Ellis, director of the Office for Protection from Research Risks (OPRR) at the National Institutes of Health (NIH), has overseen an unprecedented series of recent research shutdowns at universities, including major institutions such as the Duke University Medical Center (see *Nature* 399, 190; 1999).

The latest came last week, when the OPRR suspended some 550 federally funded researchers at the University of Alabama at Birmingham. The office has charged that the targeted universities are not living up to the government's requirements for protecting human subjects. The shutdowns have cost the institutions concerned millions of dollars.

But Ellis's supporters say he faces punishment for his activities when the OPRR is dissolved and responsibility transferred to Shalala's office. The move was recommended by a panel of advisers to Harold Varmus, then the NIH director, in a report last June (see *Nature* 399, 514; 1999).

According to Ellis's backers, top NIH

officials are using the move as an opportunity to press for the installation of a more research-friendly 'human subjects tsar' as director of the new office, to be called the 'Office for Human Research Protections'. The NIH declines to comment, except to say through a spokesman that "the search for the OPRR director is entirely in the hands of the office of the secretary".

Arthur Lawrence, the deputy assistant-secretary for health operations at HHS, says he has "no basis" for responding to the allegation. But he points out that HHS is hiring for "a completely new job", and that a wide search for qualified applicants is therefore appropriate. The job is a senior position in executive service, ranking above the current OPRR director in power and pay.

The job description posted by HHS calls for a "statesman" with "national recognition for his/her accomplishments in scientific research". Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, notes that the requirement for a statesman "might exclude someone like Gary Ellis, who is clearly doing a good job".

Ellis, 45, has a PhD from Northwestern University in biological sciences and specialized in medical reproductive endocrinology as a postdoc. He came to Washington as a science and health policy researcher in 1983, and has directed OPRR since 1993.

The OPRR has disciplined more than 10 per cent of the nation's medical schools during Ellis's tenure. The pace and severity of its sanctions grew dramatically after the HHS inspector general issued a voluminous report in June 1998 arguing that the system for ensuring protection for human subjects was "in jeopardy". Members of Congress have argued that federal watchdogs were failing to do their part to enforce protection (see *Nature* 393, 610; 1998).

Since October 1998, Ellis's office has shut down federally funded research for varying periods at seven institutions, of which Duke and the University of Illinois at Chicago are the most prominent. But researchers and institution officials have complained that the offences for which they were shut down were infractions related to procedure and paperwork, and that the punishment has been out of proportion to the offence because subjects had not been exploited.

To suspend all research at an institution is "devastating", says Robert Levine, a professor of medicine at the Yale University School of Medicine and chair of the Yale medical school committee that reviews the ethics of experiments with human subjects. Levine has been asked by HHS to apply for the new director's job. Ellis declines to comment, except to say that he is applying for the post.

Meredith Wadman

Top UK epidemiologist suspended after complaints

London

Roy Anderson, a professor of zoology at the University of Oxford and the director of the Wellcome Trust Centre for the Epidemiology of Infectious Disease, has been suspended on full pay following formal complaints from two female members of his staff.

Paul Harvey, the head of the zoology department, says that the suspension is a "neutral move while the allegations are under investigation and so do not say anything about innocence or guilt".

The suspension follows an investigation and preliminary report by university officials to the registrar and vice-chancellor, who have apparently decided there is a case to be answered.

The case centres on an appointment to a readership in epidemiology at the end of last year. The position went to Sunetra Gupta, considered by the appointment committee as the best candidate. As well as a highly regarded scientist, Gupta is perhaps better known as a novelist — her most recent book, *A Sin of Colour*, came out last year.

In early December, Gupta filed a



Anderson: allegations over remarks.

complaint which is said to concern public allegations made by Anderson on the means by which she had obtained her appointment and on her relationship with Harvey. The delivery of the preliminary report initiates procedures that look likely to lead to a private

disciplinary hearing.

Anderson is a senior figure in UK biomedical research and an expert on the epidemiology of BSE (see *Nature* 392, 533; 1998). He was head of the zoology department at Oxford during 1993–98, and has been director of the Wellcome Trust centre there since 1994.

The university issued a statement last week confirming that a director of "one of the university's research centres" had been suspended. *Nature* has been investigating under its statutory procedures.

"We can't say very much because it would be wrong to prejudice the points that will be made at the various stages of this process," says a university spokesperson. "Suspension is basically a step to mean that everyone can, as much as possible in these situations, get on with their work."

A member of the university with legal experience, and four senior members of the University's Congregation — the 'parliament of dons' that has final responsibility for all legislative matters — will chair a hearing. Anderson may attend this hearing, have representation and call and question witnesses.

The timescale for the disciplinary hearing is not yet known. But a university spokesperson said, "it is in everyone's interest it is done as quickly as possible, but there are practical issues to do with getting together the people needed".

Another researcher at the Wellcome Trust centre has complained about Anderson's comments on the appointments process. Attempts last week to contact both Anderson and Gupta were unsuccessful.

Natasha Loder

US government shuts down Pennsylvania gene therapy trials

Washington The US Food and Drug Administration (FDA) last week shut down five gene therapy trials at the University of Pennsylvania indefinitely, after finding “serious deficiencies” in the supervision and monitoring of a trial that resulted in the death of a patient. The death of 18-year-old Jesse Gelsinger, of Arizona, was the first directly attributable to gene therapy (see *Nature* **401**, 517; 1999).

The FDA identified 18 areas where the university failed to provide proper supervision and monitoring of the clinical trial to treat a liver enzyme deficiency. Two of these violations were identified during a hearing by the Recombinant DNA Advisory Committee last December. Other violations include failure to tell patients of the deaths of two monkeys who received the experimental treatment, improperly completed informed-consent forms, and patients being placed in cohorts other than the ones they were initially approved for without notifying the FDA.

James Wilson, director of the university's Institute for Human Gene Therapy (IHGT), which ran the trial, must respond to each of the points raised by the FDA, although an

FDA spokesperson declined to provide a deadline. After Wilson's reply, the FDA may elect to take further action. Judith Rodin, the University of Pennsylvania president, has appointed a committee of scientists to review IHGT's supervision of clinical trials.

Biological physics papers make it on to Medline

Paris The American Physical Society has won a long struggle to have biological physics papers from all its journals—including its prestigious *Physical Review Letters*—indexed on Medline. Repeated requests had been made by Martin Blume, the society's editor-in-chief, Margaret Foster, its editor, and two of the organization's members, Steven Schiff of George Mason University and Robert Austin of Princeton University.

“We had made a strong effort to get Medline listing because of the growing importance to biomedical researchers of papers in biological physics, and also the importance of listing to the authors of the papers,” says Blume. Medline's decision to accept the biological physics papers was “helped”, he says, by encouraging noises from Harold Varmus, the former director of the US National Institutes of Health, and Rita Colwell, director of the US National Science Foundation.

Institut Pasteur limits bosses' terms of office

Paris The Institut Pasteur in Paris is to introduce a compulsory 12-year limit for the term of office of heads of laboratories, in a bid to inject fresh blood into the institute. The measure, announced on Monday by Philippe Kourilsky, its new director, will be applied retroactively, and will therefore affect over half of the current laboratory heads.

The institute also plans to create a series of small laboratories with a lifespan of five years, to be led by young scientists. One aim of the move, said Kourilsky—who described the plan as “one of the surest means to win in excellence”—was to lure back bright young French scientists from postdoctoral positions abroad, in particular the United States.

India plans to upgrade Internet connections

New Delhi The Indian cabinet last week approved an ambitious US\$350 million scheme to establish a high-speed information highway to carry Internet traffic between universities and research centres in India and leading international scientific institutions.

The Indian government is promoting the

project, called Sankhya Vahini, with IU-Net, an associate company of Carnegie Mellon University in Pittsburgh. IU-Net will have 49 per cent equity in the venture. Scientists will be able to access the Internet at 1,000 times the speed currently available from Indian Internet service providers.

Calls for loan waiver for medics in research

Washington Citing an impending shortage of physician–scientists, the Federation of American Societies for Experimental Biology (FASEB) last week called for a programme of ‘loan forgiveness’ for US medics who choose a career in clinical research over a more lucrative life in private practice.

“Physician–scientists can be likened to an endangered species,” says FASEB president David Kaufman. Although the organization declined to cost its proposal, it favours loan forgiveness for 100 clinical MDs a year, starting within two years, with plans to eventually scale up to 400.

Report warns Germany against neglecting science

Munich Pharmaceutical and chemical research in Germany lost ground dramatically in the 1990s, according to a new report. But

thanks to continuously high investment in transport technology, and to new research efforts in biotechnology and telecommunication, the country still has an overall technological lead in Europe.

The report, *Germany’s Technological Capacity and Efficiency*, was commissioned by the federal research ministry from five economic research institutes. It concludes that “current investment in research and technology in Germany is insufficient in the long term”. In response, research minister Edelgard Bulmahn has called on industry to spend the savings from an announced tax reduction on increasing investment in research and technology.

Repaired Hubble gets frosty reception

Washington The US space agency NASA has released the first pictures from the Hubble Space Telescope following a successful mission to service it in December. These show a planetary nebula — the glowing remains of a dying, Sun-like star — that has been nicknamed ‘Eskimo’ (above right).

The picture shows a bright central region, which is a bubble of material being blown into space by the central star’s intense ‘wind’ of high-speed material. The origin of the comet-shaped features in the outer ring of



gas is less certain. This relic has been nicknamed the Eskimo nebula because, when viewed through ground-based telescopes, it resembles a face surrounded by a fur parka.

Communists hold science posts after Duma reshuffle

Moscow Ivan Mel’nikov has retained his position as head of the committee on education and science in the distribution of posts in the newly elected State Duma, the lower chamber of the Russian parliament. Yuri Maslyukov, the former deputy prime minister in Evgeny Primakov’s cabinet, has been appointed head of the committee on industrial construction and science-based technologies. Both are members of the communist faction.

Cell contamination leads to inaccurate data: we must take action now

Sir—In 1981, Nelson-Rees *et al.* found that many cell lines had been unwittingly switched or cross-contaminated with HeLa cells¹. Despite that warning, the number of published cases of cross-contamination is still increasing. Reference culture collections use techniques including DNA fingerprinting to authenticate their cell stocks^{2–4} and continue to discover HeLa contamination^{2,5}. In such cases, the only remaining characteristic of the original cultures is their name!

Misrepresentation of such cultures is perpetuated in scientific publications where the original designation is retained, including several HeLa derivatives of established value as reference cells, for example Hep2c, INT407 and KB — used respectively in virology, bacterial studies and cancer research⁵.

The American Type Culture Collection (ATCC) recently found 15 cases that were not authentic among newly acquired cell lines (see <http://www.atcc.org>). Furthermore, a recent survey of 252 human tumour cell lines, performed by the German Collection of Microorganisms and Cell Cultures (DSMZ), identified 45 (18%) that had been cross-contaminated by their originators⁶.

It is high time that the true origins of 'cross-contaminated' cultures are acknowledged, for transparency and to raise awareness amongst those new to cell culture. Action at several levels is required.

We recommend clear identification of cross-contaminated cultures in catalogue entries of culture collections. Thus, we propose that HeLa-contaminated cultures should carry the appendage '(HeLa)' in addition to the official cell name.

Resource centres and scientific organizations (such as the World Health Organisation international cell banks) should collaborate to authenticate cell stocks and ensure unrestricted access for research.

When new cell lines are established, representative samples of the original tissue, cells or DNA should be archived by originators for later authentication of cell stocks.

Authentication of new cell lines should be a prerequisite for publication. We strongly recommend that this should be by submission to a culture collection (not necessarily for immediate release to other investigators). This protects the intellectual investment of the originator and is a prerequisite for certain patents.

Cell lines should be disseminated only if they are from authenticated sources, such as bona fide culture collections; recipients should refer to guidelines on best practice

in cell culture, such as those produced by the UK Co-ordinating Committee on Cancer Research and the US Food and Drug Administration^{7,8}.

Proper documentation must accompany a cell line being transferred between laboratories, including a description of the cell line, its origin and provenance, confirmation of authenticity and freedom from mycoplasma, and biohazard information.

Such precautions could save months or years of wasted work. Those using cell culture in research and development should remember a phrase coined in information technology: "garbage in, garbage out".

G. N. Stacey

National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Herts EN6 3QG, UK

Other signatories of this letter:

J. R. W. Masters *Institute of Urology, University College London,*

67 Riding House Street, London W1P 7PN, UK

R. J. Hay *American Type Culture Collection, 10801 University*

Boulevard, Manassas, Virginia 20110-2209, USA

H. G. Drexler, R. A. F. MacLeod *German Collection of Microorganisms*

and Cell Cultures, Mascheroderweg 1b, 38124 Braunschweig, Germany

R. I. Freshney *CRC Department of Medical Oncology, University of*

Glasgow, Garscube Estate, Bearsden, Glasgow G61 1BD, UK

1. Nelson-Rees, W. A., Daniels, D. W. & Flandermeyer, R. R. *Science* **212**, 446–452 (1981).

2. Gilbert, D. A. *et al. Am. J. Hum. Gen.* **47**, 499–517 (1990).

3. Gignac, S. M. *et al. Leukaemia Lymphoma* **10**, 359–368 (1993).

4. Stacey, G. N., Bolton, B. J. & Doyle, A. *Nature* **357**, 261–262 (1992).

5. Bolton, B. J., Stacey, G. N. & Doyle, A. *In Vitro* **29A** (part II), 123A (1993).

6. MacLeod, R. A. F. *et al. Int. J. Cancer* **83**, 555–563 (1999).

7. Working Party for United Kingdom Co-ordinating Committee on Cancer Research. *UKCCCR Guidelines for the Use of Cell Lines in Cancer Research*. (UKCCCR, London, 1999).

8. *Points to Consider in the Characterisation of Cell Lines used to Produce Biologicals* (Office of Biologics Research Review, FDA, Bethesda, Maryland, 1997).

Mitchell saw the new vista, if not the details

Sir—For a researcher in the area of bioenergetics, like myself, it was a delight to read Leslie Orgel's Millennium Essay¹. Peter Mitchell's formulation in 1961 of the chemiosmotic theory as an alternative to the covalent intermediate opened a completely new vista in the field.

But Mitchell never suggested that the electron-transfer complexes in respiration and photosynthesis function as proton pumps; in fact, he was an ardent opponent of the idea of redox-linked proton pumps. For example, after Wikström reported² that cytochrome oxidase is a proton pump, Mitchell wrote an article called "Cytochrome *c* oxidase is not a proton pump"³. Mitchell's idea of how the proton gradient across the membrane is created was the 'redox loop'. According to this concept, a hydrogen atom is first transferred from one side of the membrane to the other, where it is split into a proton and an electron; the electron is then transported back across the membrane, leaving the

proton behind. The difficulty with this clever mechanism is that it can only have a H^+/e^- stoichiometry of -1 , which would mean that about half of the energy available from the electron-transfer reactions would be wasted.

The Royal Swedish Academy of Sciences awarded Peter Mitchell the Nobel Prize for Chemistry in 1978, on the recommendation of its Nobel Committee for Chemistry, of which I happened to be chairman at the time. The committee was well aware of the fact that Mitchell was wrong about how the proton gradient is created (proton pumps rather than redox loops) and how it is utilized to make ATP (conformational coupling rather than mass action). This is why the academy's citation read rather vaguely "for the contribution to the understanding of energy transfer through the formulation of the chemiosmotic theory".

As Mitchell was right only on the phenomenological and not on the mechanistic level, one could argue that he should have been given the Nobel Prize for Physiology or Medicine instead. As one of my teachers used to say: "Physiology is that part of biochemistry which we do not yet understand".

Bo G. Malmström

Department of Chemistry (Biochemistry and Biophysics), Göteborg University, PO Box 462, SE-405 30 Göteborg, Sweden

1. Orgel, L. E. *Nature* **402**, 17 (1999).

2. Wikström, M. *Nature* **266**, 271–273 (1977).

3. Mitchell, P. *FEBS Lett.* **88**, 268–272 (1978).

Give credit where it's due (not to me, this time)

Sir—In his interesting review of Stan Prusiner's *Prion Biology and Diseases*¹, Colin Masters inadvertently gave me credit for showing "that accessory chaperones... may facilitate the conversion of a normal host-cell protein into a pathogen". Although I can take credit for discovering that [URE3] and [PSI] are prions of yeast, the credit for discovering the role of chaperones in this phenomenon goes to Yuri Chernoff and colleagues^{2–4}.

Any chagrin I may feel at being inappropriately credited is measurably mitigated by the knowledge that there will inevitably be occasions when I will be 'unintentionally' ignored.

Reed B. Wickner

Laboratory of Biochemistry and Genetics, NIDDK, National Institutes of Health, Bethesda, Maryland 20892-0830, USA

1. Masters, C. L. *Nature* **402**, 581–582 (1999).

2. Chernoff, Y. O. & Ono, B.-I. in *Protein Synthesis and Targeting in Yeast* (eds Brown, A. J. P., Tuite, M. F. & McCarthy, J. E. G.) 101–107 (Springer, New York, Heidelberg, 1992).

3. Chernoff, Y. O., Lindquist, S. L., Ono, B.-I., Inge-Vechtomov, S. G. & Lieberman, S. W. *Science* **268**, 880–884 (1995).

4. Newnam, G. P., Wegrzyn, R. D., Lindquist, S. L. & Chernoff, Y. O. *Mol. Cell. Biol.* **19**, 1325–1333 (1999).

Science moves to centre stage

Science is playing an increasing part in many decisions made by the UK Parliament, and probably in other legislation. What are the implications?

Ana Padilla and Ian Gibson, MP

We have undertaken an analysis of questions, motions and debates in the British Parliament. The proportion related to science and technology has increased about sixfold over the past decade, with those concerned with biological and environmental sciences accounting for most of the growth. This increase indicates that scientific issues are playing an increasing role in political decision-making.

Parliamentarians with an interest in issues related to science and technology can use various channels to help focus political attention on them. These range from direct scrutiny of ministers by parliamentary questioning to the production of exhaustive research papers and reports on particular subjects. The latter is normally carried out by select committees, the Parliamentary Office of Science and Technology and the House of Commons Library, as summarized in Box 1.

To assess trends in science and technology-related issues brought before Parliament, we analysed parliamentary questions in the House of Commons and House of Lords, Early Day Motions, debates and Commons' select committees' reports (see Box 1). The source material for parliamentary questions was provided by a digest published by the Parliamentary and Scientific Committee in its journal, *Science in Parliament*. This includes all science and technology-related parliamentary questions (written, oral and supplementary) that have been asked in the Houses of Commons and Lords.

Parliamentary questions

We analysed a total of 14,459 science and technology-related parliamentary questions asked during the past ten years, classifying them into four categories: physical sciences (information technology, physics, engineering); life sciences (medicine and food); environment (including energy); and science policy (for example, funding). The total number of science-related questions per parliamentary session was then normalized to the total number of questions asked in both houses for the same period of time, to establish a consistent basis for analysis.

The results show (see Fig. 1, overleaf) that the percentage of questions related to science and technology has risen over the past ten years from less than 1% in 1988–89 to around 6% in 1998–99. This has not been a steady rise; several fluctuations are noticeable. For instance, there is a marked drop in

Good science does not necessarily present an obvious policy pathway.

election years, 1992–93 and 1997–98. On the face of it, this is surprising, as UK election campaigns tend to comprise a relatively short period (just over a few weeks). But both of these elections were predictable, coming at the end of the five-year maximum term. The drop probably reflects the change in parliamentary focus in the months leading up to a general election, when there is less

concern for strategic issues and more political jostling for position.

Looking at specific issues (see Fig. 2 on page 359), the bulk of the parliamentary interest in science has been dominated by the life sciences (medicine and food) and environment (including energy), which together account for more than 70% of all science and technology-related questions. Throughout the period, however, the relative positions of life sciences and environment have changed. There were more environmental questions between the 1988–89 and the 1994–95 parliamentary sessions, reflecting increasing interest in environmental issues in the run-up to, and immediate aftermath of, the United Nations Earth Summit in Rio de Janeiro. The

Box 1: Science information channels in Parliament

● **Parliamentary questions**, either written or oral, act as a direct channel for the scrutiny of ministers.

● **Early Day Motion**, or EDM, is the term used to describe notices of motions given by Members of Parliament that are not generally expected to be debated. The tabling of an EDM is a device to draw attention to an issue, and to elicit support for it by inviting other members to add their signatures to the motion. As a general rule, any EDM supported by more than 100 signatures is taken seriously and brought to the minister's attention, playing a part in the general debate.

● **Debates** in the House of Commons are raised through a ballot procedure, as well as following ministerial statements. Debates are also raised sometimes following select committees' reports (for example, the Dearing report on higher-education funding). They tend to be between 30 and 90 minutes long.

● **Select committees** are committees of inquiry, which mostly proceed by taking evidence and making reports on their findings. Select

committees are appointed at the start of a new government, for a whole Parliament or session, and their members mirror the proportional partisan representation in Parliament. The chairman of each committee is chosen by the party whips. Select committees are serviced by the clerks' office and usually shadow departments, either examining their work or taking specific issues as subjects for inquiries. Committees have disappeared or been fused to other committees, to achieve a better coverage of events.

● **The Parliamentary Office of Science and Technology (POST)** operates as an advisory body for both houses, producing information briefings and wider reports on a range of science and technology-related topics, from life sciences to the environment and physics. The topics covered by POST are decided by its board members. Parallel to this, POST advises select committees on scientific and technological aspects of particular inquiries.

● **House of Commons Library, Science and Environment Section.** Parliamentarians also

have access to library facilities, including the work carried out by this section, whose objectives include the provision to individual members and their staff of information appropriate to their particular needs. The library also produces briefings on topics of immediate interest, often related to legislation as it is passing through the House of Commons.

● **The Parliamentary and Scientific Committee (PSC)** was established in 1939 as a body associated with but not part of Parliament, and is an important forum for the exchange of information and ideas between the worlds of politics and of science, technology and engineering. The PSC also edits the journal *Science in Parliament*, which publishes digests of all science and technology-related parliamentary questions (both written and oral, as well as supplementary), that have been asked in the House of Commons and the House of Lords. These questions have been selected by PSC staff and subjected to internal quality controls. As such, we assume that this classification has been accurate throughout the period analysed.

prevalence of environmental issues has declined since then, while questions related to life sciences have steadily increased since 1991–92 and now dominate. Several factors can account for this, such as recurrent food scares and a widespread awareness of health issues, for example the scandal over bovine spongiform encephalopathy (BSE) (1996–97 session) and the debate on genetically modified (GM) foods (1998–99 session).

Questions on the physical sciences and science policy have, on the whole, not changed dramatically during the ten-year period analysed. In the case of physical sciences this might seem surprising, in the light of the radical increase in the development and use of information and communication technologies. However, these issues seldom constitute a direct ‘threat’ to the public or the environment, and hence there is less political incentive to enter into debates. This may also reflect the fact that some Members of Parliament (MPs) are out of touch with the development of the web and the use of electronic mailing facilities. This is likely to change, particularly with the current government’s enthusiasm for e-commerce and electronic service delivery. However, we found no single issue dominating the parliamentary questions in any session, and cannot therefore account for the increase within particular categories. The overall trend in all such questions must therefore reflect a general increase in a wide range of subjects concerning science and technology.

Early Day Motions

We also analysed Early Day Motions for the past five parliamentary sessions (a period for which data were readily available), classifying them by the number of MPs supporting them. As stated in Box 1, Early Day Motions carrying more than 100 signatures are considered to be ‘important’. Our results show that the percentage of important motions related to science and technology has nearly doubled, despite a significant decline at the end of the previous government’s administration. This is also reflected in the total number of important motions, which dropped from 122 during the 1994–95 session to 46 during the 1996–97 session. Since the current administration came to power in 1997, the number of important motions has increased again, reaching 104

in the 1998–99 session, with science and technology-related subjects accounting for nearly 60% of the total. Although Early Day Motions might not be the best indicator for an analysis of science and technology within Parliament, they follow a similar pattern to that observed for parliamentary questions.

Debates and select committees

Another important way for MPs to raise a particular issue is through debates. An analysis of all 4,350 debates held at the House of Commons from the 1994–95 session to the 1998–99 session (June 1999 inclusive) found that 16% of the total number of debates concerned science and technology, again dominated by life sciences (42%, after excluding all debates about the National Health Service that did not concern science and technology) and the environment (38%).

Virtually all select committees have dealt with some aspect of science and technology during the past ten years, even those that do not usually encounter such topics — examples include the Social Security Select Committee (data matching; year 2000); the Welsh Affairs Select Committee (wind farms); and the Home Affairs Select Committee (electronic tagging for criminals). However, we do not think it is worthwhile to look for trends in select-committee inquiries for two main reasons. First, there is a limit to the number of inquiries each committee can conduct. And second, most committees are tied to scrutinizing particular government departments, and so their own agendas reflect government action.

Increasing importance of science

Our analysis shows that scientific issues have become increasingly important in the work of the UK Parliament over the past ten years. But does this reflect an increasing importance of science and technology within Parliament, or merely an increasing sophistication and efficiency of lobby groups at getting these issues on the agenda? Our guess is that both are factors. However, we feel that at least part of the apparent rise reflects the growing importance of science and technology for regulatory legislation, exemplified by issues such as the BSE crisis and calls for ‘sound’ science during the attempts to regulate GM foods, as well as a recognition that

the best medical practice is evidence-based.

We think the trend will certainly continue. Just last November, the Queen’s Speech signalled the government’s intention to introduce during the new parliamentary session 28 pieces of legislation in which issues related to science and technology feature widely. These include bills to promote electronic commerce and electronic government, modernizing utility regulation, encouraging skills and training for young people, introducing mandatory drug testing in the criminal justice system, allowing for the interception of communications, protecting wildlife, increasing transport safety, allowing charging to reduce road-traffic congestion, facilitating freedom of information and encouraging more efficient water use. In addition, significant scientific issues for political debate, for example the Human Genome Project and the subsequent gene-patenting issues, as well as trade-related aspects of intellectual property rights, are waiting just around the corner.

It would be interesting to know whether the apparent upward trend occurs in other developed countries. Although we did not collect data on other national legislatures’ dealings with issues involving science and technology, there is circumstantial and anecdotal evidence. For instance, many of the issues to be dealt with by the UK Parliament are also of global relevance, often concerning international trade and standards. Moral, legal and ethical questions will be raised everywhere regarding several key issues: the Human Genome Project relating to the ownership of genetic sequences, genetic testing and eugenics; the production of GM organisms, either for food or for human therapeutics; sustainability of health-care systems in the face of demographic changes and increasing demand; climate change and the need to reduce emissions of greenhouse gases; the introduction of innovative technologies to minimize the creation and impact of industrial waste; and further growth of the Internet, including issues related to security, money laundering and privacy.

Hence, there is a need for an infrastructure for science and technology in the national legislature of each country, involving parliamentary committees as well as non-partisan analysis and information organizations. One measure of the growing importance of science and technology is the increase in the number of these non-partisan institutions from one in 1972 to 15 now — for example, the Danish Board of Technology, the Rathenau Institute in the Netherlands and the Bureau for Technology Assessment of the German Bundestag. All are European and most are cooperating within the European Parliamentary Technology Assessment network. Several other countries, for example Korea, Japan and Canada, are considering setting up similar bodies.

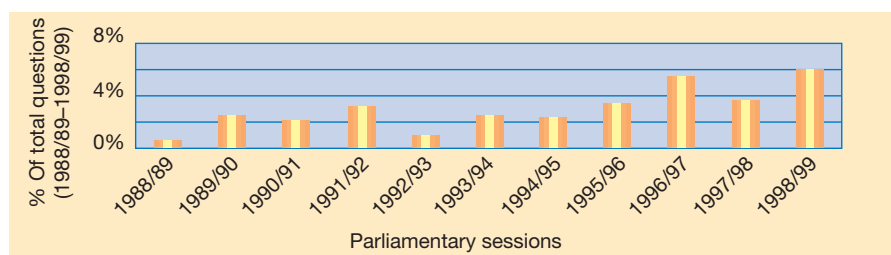


Figure 1 Science and technology in Parliament (1989–99) represented as a percentage of total parliamentary questions.

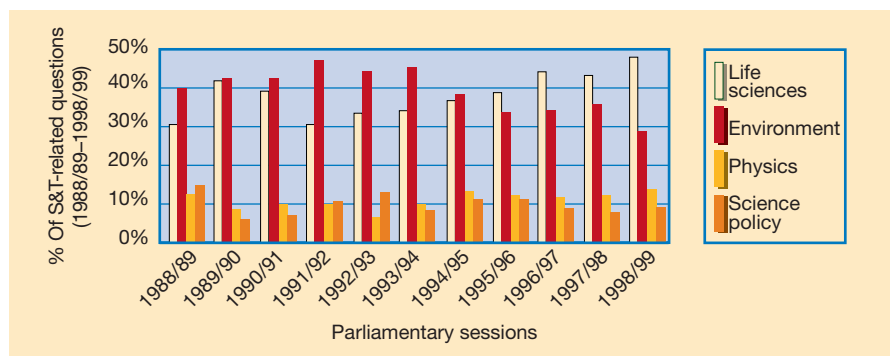


Figure 2 Science and technology-related parliamentary questions by category (1988–89 to 1998–99 sessions).

International communication is also improving between environmental groups, consumer-interest groups and other pressure groups. A successful campaigning group will advise other groups on productive approaches (for example, North American environmental groups have been seeking advice from Friends of the Earth and Greenpeace on the running of anti-GM food campaigns). The increased availability of information on the Internet turns international lobbying into a significant reality. In other cases, global scientific issues will have differential regional impacts (for example, population growth, the AIDS epidemic in Africa, biodiversity or sustainable development).

Open debate

If, as we argue above, science and technology issues are set to account for an increasing proportion of parliamentary business in legislatures around the world, what are the likely implications? Parliamentarians are expected to understand and reflect the concerns and values of those they represent, and to ensure measures are taken to maximize welfare and minimize risk. Clearly, this applies to all aspects of politics, but in science and technology the complexity and pace of change are such that Members often struggle to fully understand them and hence to grasp their wider social repercussions. So it is vital that parliamentarians obtain and use unbiased information on technical matters and on people's concerns.

Parliament is a forum where questions can be raised and voices expressed. Parliament's strength lies in its ability to guarantee media scrutiny, to activate lobbyists and pressure groups and to open up the potential for legislation and public debate. Parliaments are independent of, but influenced by, governments, pressure groups and business, and hence provide the ideal setting for the open and inclusive debate necessary to resolve many issues involving science and technology. As a body that interprets the views of the public, Parliament must deal with the wider scepticism towards science and technology that has been generated in part by commercial involvement. Among

other factors is the loss of trust in institutions that manage risk on behalf of the public (for example, food safety).

Members of Parliament are increasingly raising scientific issues with little input from extra-parliamentary sources. In many cases, science dominates discussions, for instance through evidence taken before the Science and Technology Select Committee, or via parliamentary response to reports from government advisory bodies. Parliament is a forum in which ministerial decisions and committee recommendations can be challenged, and the debate broadened. Nevertheless, an underlying problem for any scientific issue is the requirement for a general understanding, as well as the fact that science rarely provides clear-cut answers. Therefore, the provision of unbiased scientific information, including an honest admission of the uncertainties involved, will be recognized as an essential tool in the process of scrutinizing governments.

It was a welcome step when Britain's prime minister said in answer to a parliamentary question that 'good' science should drive the debate on GM organisms. But after the experiences of 'good' science in the BSE debate, when there was political suppression of data, this criterion is definitely not enough on its own. Indeed, 'good' science is also required to provide economic benefits for society. It is not simply defined by peer review and publication. Demands are placed on the United Kingdom, as on any country, to stay ahead in matters of science and technology and medicine, and to support the view that the public will benefit directly from the knowledge-based economy, and that it should receive the best advice on matters concerned with food safety, drug development, and so on. The recent Bill presented by one of us (I.G.) calls for a doubling of the science budget which would be essential for scientific advance and support over the next ten years¹. The enterprise culture requires underpinning by science and technology.

Scientific knowledge is playing an increasing part in political decision-making. Scientists themselves will have to recognize that blind public acceptance of their work

cannot be taken for granted. As a consequence, they and their representative bodies will have to examine their roles *per se* and in unfamiliar territory, both political and public. The media as well as the scientific press will become an essential place for proactive discussion, debate and presentation.

Similarly, parliamentarians must be more willing to adopt procedures for receiving direct input from the general public, rather than relying on traditional consultation methods, as an additional source of information. Examples of such an approach already exist in Denmark and the Netherlands.

We need to find an interactive mechanism by which all stakeholders can participate in science and technology-related issues. There is a need for a debate on the extent to which governments and parliaments should seek or take account of public opinion, the point at which this should occur and the mechanisms provided for such a task². The debate on how this should be taken forward in the United Kingdom has only just started and it is too early to point to any one model. However, a key requirement is that the mechanism should be embodied in an official institution that is 'plugged in' to the policy process.

At the same time, it is essential that we understand what science can offer in terms of certainty and risk. Science is about probabilities and uncertainty. Although it is possible to prevent bad decisions being made on the basis of bad science, good science does not necessarily present an obvious policy pathway. The role of science is essential but not necessarily definitive. In contrast, the public has strong demands to be heard, and wants answers to its questions. In particular, many people react fiercely against risk analyses that artificially compare unrelated dangers^{3,4}. A new compact, then, is needed between parliamentarians, scientists and the general public that will involve new, as yet unknown, institutional arrangements. Let the experiments begin.

Ana Padilla can be contacted through the Parliamentary Office of Science and Technology, 7 Millbank, London SW1P 3JA, UK.
e-mail: padillaa@parliament.uk
Ian Gibson, MP is at the House of Commons, London SW1A 0AA, UK.

1. Gibson, I. *National Science Strategy* (House of Commons Bill 99/48; 1999).
2. ESRC *The Politics of GM Food: Risk, Science & Public Trust* (Economic & Social Research Council, Special Briefing No. 5, October 1999).
3. Gibson, I. *A Safer Attitude to Risk* (Science & Public Affairs, 1999).
4. POST *Safety in Numbers? Risk Assessment in Environmental Protection* (Parliamentary Office of Science and Technology, 1996).

Acknowledgements We thank Annabel Lloyd at the Parliamentary and Scientific Committee; the Science and Environment Section at the House of Commons Library; the journal office at the House of Commons; and the clerks at the select committees for providing the raw data for the analysis. Special thanks go to Peter Border, David Cope and Gary Kass at POST for constructive criticism and support. A.P. is research assistant to I.G.

Instincts for the past tense

Does the structure of the English language reflect the structure of the mind?

Words and Rules: The Ingredients of Language

by Steven Pinker

Weidenfeld & Nicolson/Perseus: 1999.

348 pp. £14.99, \$26

David Poeppel

Have you ever wondered why the past tense of sing is sang but the past of fling is flung while that of bring is brought? And, more importantly, why a reviewer can ding your grant, and why you can complain that a reviewer dinged it — but not that a reviewer dang, dung or dought it? If you are curious about these sorts of language phenomena (and many others), get Steven Pinker's *Words and Rules*. Finding a reader-friendly balance between humour, irreverence and analytically retentive scholarship, Pinker unpacks a remarkable variety of facts associated with the distinction between regular and irregular English words and their structure. If you like (English) words and want to know why they are put together in the way they are, this book is definitely for you.

If you have not been too worried about the English past tense, is the book still worth reading? It is, if you are willing to entertain the argument that the detailed theoretical-linguistic, psycholinguistic and neuro-linguistic dissection of the English past-tense system sheds light on a much larger issue: the structure of the human mind. The position developed in *Words and Rules* is that the small and circumscribed linguistic system under consideration should be viewed as a model system for the scientific study of language. Biologists focus on model organisms such as *Drosophila*, zebrafish or *C. elegans*, and physicists develop theories based on carefully restricted model systems such as inclined planes or coupled oscillators. Pinker suggests that the English past tense can serve as the “*drosophila* of psycholinguistics” — as a model system for investigating the structure of language and mind from every possible perspective.

The intellectual foundation providing the backdrop for the research is the very old but still vibrant debate between rationalists and empiricists, a debate based on ideas developed by René Descartes, Thomas Hobbes and Gottfried Leibniz and their counterparts John Locke and David Hume. Is the mind a system of ‘reckoning’ using abstract representations (symbolic computation, in contemporary terms), or is it an associative memory in which computation builds on the frequency and similarity among remembered items (a connectionist architecture, in contemporary terms)? Are

cognitive systems such as language largely innately specified and richly structured, or are they acquired on the basis of general learning principles?

In the domain of language, these issues have been explored in many contexts, including the formation of the past tense in English. The past-tense theory that forms the basis for much current research was proposed by Noam Chomsky and Morris Halle in 1968. Working in the framework of rule-based generative linguistics, they showed that the entire repertoire of English irregular verbs (there are approximately 165, compared with thousands of regular verbs) could be accounted for by merely three phonological rules. In 1986, an influential paper by David Rumelhart and James McClelland showed that a relatively simple connectionist network could also capture a surprisingly large set of facts about past-tense formation without appealing to a modularized architecture and rules.

Pinker explains both proposals, and their strengths and weaknesses, in some detail. The discussion of the Rumelhart–

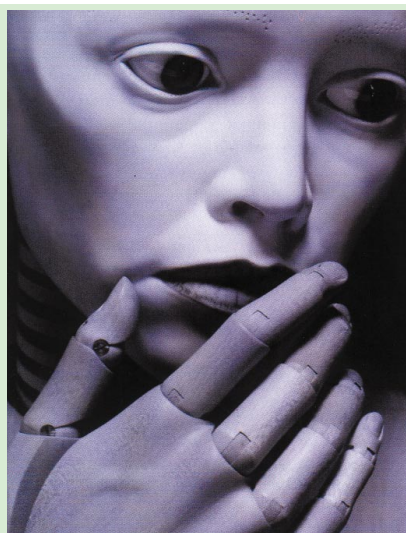
McClelland model is particularly clear, and enables the reader to reason through each step relevant to the function of this connectionist network.

Pinker attempts to unify the two views. The approach is based on the past dozen years or so of work in his lab, with crucial contributions made by the cognitive scientists John Kim, Gary Marcus, Sandeep Prasada, Alan Prince, Michael Ullman, Fei Xu, and others. The essence of the proposal is, unsurprisingly, that you can have your cake and eat it too.

Pinker argues that the Cartesian–Hobbesian ‘reckoning’ approach worked out by Chomsky and Halle is the right way to think about regular inflection, but the Humean–Lockean associationism approach developed by Rumelhart and McClelland is right for irregular inflection.

Specifically, the mind is made up of a “frequency- and similarity-loving associative memory and a promiscuous combinatorial grammar”. This hybrid model is argued to be preferable because it captures the sensitivity of speakers to family resemblances among irregulars (for example bind–bound, find–found) as well as the algebra-like computation of regular word forms. Importantly, both mechanisms are active all the time. Whenever a speaker or listener encounters an irregular item, the responsible mechanism is associative memory, which immediately supplies the relevant (memorized) past-tense form. In contrast, for regular verbs no memorized past-tense forms exist and the correct form is thus computed online, by adding -ed.

Pursuant to the “past-tense-as-psycholinguistic-zebrafish” approach, Pinker goes on to discuss the model in action. Chapters on language processing, language acquisition, and the German regular–irregular distinction cover the range of effects with which the model deals effectively. These chapters are fun to read, and reflect the fact that the group has worked on these problems for many years. The two chapters with the broadest implications discuss the possible neurological basis of the regular–irregular distinction and the connection of regulars and irregulars to the nature of categories in cognition. The neurological work makes connections between memory systems (and other cognitive systems) and regularity in language. In particular, neuropsychological and neuroimaging data suggest that irregulars are associated with a temporal-lobe network implicated in memory. The computation of regulars, on the other hand, is associated with a frontal network



A world of silent communication

Pupil: Pose 1, by Katherine Wetzel and Elizabeth King, is one of many images to be found in *Ghost in the Shell: Photography and the Human Soul, 1850–2000* (Los Angeles County Museum of Art/MIT Press, \$59.95, £37.50) by Robert A. Sobieszek. The book traces the history of photographic portraiture, and analyses the premise that the human face expresses the essence of the human character.

implicated in the execution of rule-like behaviour.

In a final chapter, Pinker argues that the regular-irregular distinction exemplifies a more general feature of the human mind. Cognition, he argues, is supported by two types of categories — classical categories (for example, odd number) and fuzzy, family-resemblance categories (for example, chair). The nature of the categories supporting cognition is an old chestnut in cognitive science, and Pinker wants to connect the psycholinguistic research to more general properties of cognitive systems. Regular words are examples of classical categories, whereas irregular words are examples of family-resemblance categories. Symbol combination is crucial for the computation of regulars, as well as classical categories more generally; and associative memory is the core mechanism subserving the representation of family-resemblance categories in general, and irregulars in particular. Insofar as both classical, definitional information and family-resemblance information constitute crucial aspects of what the mind must deal with, both mental mechanisms are integral requirements for a functioning cognitive system.

For a broad audience, *The Language Instinct* (Penguin, 1995) was Pinker's more amusing book, and *How the Mind Works* (Penguin, 1998) was more speculative. *Words and Rules* is an academic work. I very much like the thoroughness and clarity of this book, notwithstanding at least one important sin of omission. Pinker thoughtfully argues for abstract representation and symbolic computation. In fact, a central concern throughout is to show that one cannot get away without a highly structured linguistic computational system. Surprisingly, Pinker does not raise what many linguists and psycholinguists consider the most interesting argument for a richly structured computational system (language faculty) that makes use of symbolic rules. It is called the poverty-of-the-stimulus argument, and it asserts that every speaker comes to know abstract properties of his or her language in the absence of any input that could provide the necessary evidence. In a work of such broad scope, the most powerful argument for structure deserves mention.

This complaint, however, is somewhat cosmetic. More importantly, the book provides a scholarly, persuasive, enjoyable and eminently readable account of important language phenomena. ■

David Poeppel is in the Cognitive Neuroscience of Language Laboratory, University of Maryland, College Park, Maryland 20742, USA.

More on language

How the Brain Evolved Language

by Donald Loritz

Oxford University Press, £35, \$45

Why not knot right?

The 85 Ways to Tie a Tie: The Science and Aesthetics of Tie Knots

by Thomas Fink and Yong Mao

Fourth Estate: 1999. 144 pp. £10

Gregory Buck

There is a little irony in the fact that I am reviewing this book. I am a modern American mathematician, well-schooled in the sartorial traditions of my field, and so would perhaps be a natural reviewer for a book entitled *The Well-Wrinkled Tee Shirt* or, perhaps, *Wearing Sandals in the Snow*. However, I teach at a liberal arts college, and so can wear a tie while teaching when I want to without risking my mathematical reputation — of course, for conferences I pull my clothes out of the bottom of the dirty-laundry pile like everyone else. (My colleagues in the economics department scoff at my tie-wearing, considering it too infrequent to be taken seriously, but they are extremists — I am pretty sure they wear ties with their pyjamas.) I have always liked tying ties, but, despite the fact that I study knot theory, like most people my tie-knot knowledge was cultural and accidental. I knew a couple of tie knots but not their names, nor could I recall where or when I learned them.

This wonderful little book by Thomas Fink and Yong Mao has changed my life. Now, when I tie a tie, I know what I am doing, and why. Fink and Mao have performed a great service for civilization, doing for tie-knot tying what Isaac Newton did for the motion of the heavens: lifting it from the darkness of secrecy, ritual and superstition to the light of rational, scientific good taste.

To accomplish this remarkable feat, Fink and Mao have employed the analytical tools of topological (and geometric) knot theory and statistical mechanics with cleverness and dexterity — introducing just enough of each to get the job done. That may sound ambitious, but this is a book aimed at the general reader. A beautifully concise, four-page appendix contains the only mathematics that could be considered challenging. The illustrations are superb — I tried nearly all the knots illustrated and got them right first time. The notation for the knots is elegant and easy to master.

The scientific force of the work is that Fink and Mao have created a formal model that captures the salient characteristics of tie-knot tying in the real world, and have then analysed the formal model, guided by the scientific lights of simplicity and symmetry, and have solved the problem completely, identifying the 85 ways to tie a tie



Knotty problem: imagine how many tie knots are tied in a day.

(given natural constraints). Their model predicts the knots most commonly used, and provides several new possibilities.

Fink and Mao have obeyed the imperative of the scientific entrepreneur: create a niche, and then fill it completely. This book is now the definitive work on tie knots, and as such is the definitive work on one of the most common applications of knot theory (and therefore of topology). The applications of knot theory are legion: a test tube of DNA may contain billions of knots, but sometimes they are hard to see. Polymers in general may gain many of their characteristics from tangling, knotting and linking, but this may not be apparent when you are holding the material in your hand. Magnetic field lines are often knotted, linked or otherwise entangled, but one doesn't often observe this on the way to the market. But now imagine the morning dressing routines around the world — imagine how many tie knots are tied in a day.

Finally, we must consider the stylistic force of the work. Fink and Mao provide an informative history of tie-knot evolution. They also provide much more — a guide to taste in knot tying. An attentive reader will learn which knot works best with a given tie and collar, and will learn tie knots that can be enjoyed as things of beauty in and of themselves (for me it was the Plattsburgh). Fink and Mao have shown that it is possible to be both smart and smart — in brains and style. And so here is a prediction: anyone who wears a tie, who is at all of a scientific bent, will enjoy this book very much. ■

Gregory Buck is in the Department of Mathematics, Saint Anselm College, Manchester, New Hampshire 03102, USA.

Also available

Ideal Knots

edited by A. Stasiak, V. Katritch & L. H. Kauffman

World Scientific, \$55, £34

Powering the planet

From Space to Earth: The Story of Solar Electricity

by John Perlin

Aatec: 1999. 224 pp. \$32, £22.50

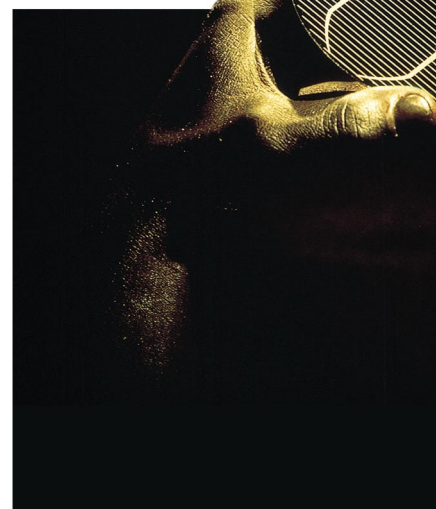
Michael Grätzel

Ever since the French scientist Henri Becquerel discovered the photoelectric effect in 1839, researchers and engineers have been infatuated with the idea of converting light into electric power. There is a magical aspect to solar cells that attracts a wide spectrum of people ranging from illiterate amateurs to highly trained professionals, from penniless entrepreneurs to wealthy sponsors. Their common dream is to collect the energy that is freely available from sunlight and turn it into the valuable and strategically important asset that is electric power.

The development of solar cells is also promoted by the increasing public awareness that the Earth's oil reserves will run out this century. As the planet's energy needs will at least double within the next 50 years, the stage is set for a major energy shortage unless renewable energy can cover the large deficit that fossil fuels can no longer furnish. Public concern has heightened recently as a result of the disastrous environmental pollution from oil spills and the frightening climatic consequences of global warming from the combustion of fossil fuels.

Fortunately, the supply of energy from the Sun to the Earth is gigantic — 32×10^{24} joules per year, or about 10,000 times more than mankind's current consumption. In other words, covering only 0.1 per cent of the Earth's surface with solar cells possessing an efficiency of 10 per cent would satisfy our present needs.

To tap into the Sun's huge energy



Light work: solar cells are becoming efficient enough to rival conventional solid-state cells.

reservoir remains, nevertheless, a major challenge for mankind. John Perlin's book gives a taste of the tremendous difficulties that early pioneers had to overcome in order to turn Charles Fritt's 1885 invention of a selenium-based solar module into today's booming photovoltaic business. Solar cells functioned very poorly for long after their discovery, with conversion yields of below one per cent. They remained a curiosity until 1954, when Calvin Fuller and Gerald Pearson from Bell Laboratories realized the first silicon-based p-n junction device, which achieved an efficiency of close to six per cent.

Despite such an impressive advance, this would probably not have been exploited commercially without the advent of the space age. It was the need to power satellites that gave the breakthrough to solar cells. Only space applications were able to afford their very high price, that is, several hundred dollars per watt of electric power produced by the solar panels. When the price of solar cells dropped to \$20 per watt they became competitive with primary batteries for remote-site terrestrial applications, mainly in the telecommunications field. Today the cost has come down to \$5 per watt, opening up commercial opportunities for building integrated solar installations connected to the grid.

Starting with the space application, Perlin gives a vivid and fascinating historical account of the advances of photovoltaics on Earth. His book tells us the success story of the pioneers of solar cells, crediting them for their imagination and perseverance. One of the remarkable individuals described is Father Bernard Verspieren, who saved many lives by installing solar-driven water pumps in Mali. Presenting the history of the development of photovoltaic cells in such a personalized manner makes it a much more lively and interesting read than a mere technical account would have done.

The book focuses strongly on silicon; gallium arsenide cells, currently the most efficient solar-light-energy converters and widely used in space applications, are not mentioned. Nor are thin-film molecular photovoltaic systems such as the nanocrystalline-injection solar cell, which has an efficiency of more than 10 per cent, becoming a credible rival to conventional solid-state cells.

I would have liked a more extensive assessment of the future potential of photovoltaics, going beyond its use as a power source for the telecommunications industry. The crucial question is whether solar cells can supply energy for the planet on a major scale. The world currently needs 10,000 gigawatts of power, a staggering figure compared with the present yearly installations of 100 megawatts' worth of photovoltaic panels.

Novel concepts that do not rely on conventional p-n junctions may well be necessary to achieve the significant cost reduction

required for large-scale applications of solar cells. The newly emerging field of molecular photovoltaics holds great promise in this respect. These systems operate like a plant's photosynthesis, mimicking a process that has worked well for the past 3.5 billion years, providing all the world's fossil energy reserves. The use of artificial photosynthesis for converting and storing solar energy provides exciting prospects for the new millennium. ■

Michael Grätzel is at the Institut de Photonique et Interfaces, École Polytechnique Fédérale, Ecublens, CH-1015 Lausanne, Switzerland.

An evolutionary odyssey

The Human Career: Human Biological and Cultural Origins (second edition)

by Richard G. Klein

University of Chicago Press: 1999. 810 pp.

\$45, £31.50

Jean-Jacques Hublin

Four and a half million years ago, the first hominids were just another species of African ape competing for new ecological niches. Like all other species before them, the pre-humans followed the well-trodden path of biological adaptation. However, in the course of their evolution, humans would differentiate dramatically from any other beings by developing increasingly complex technologies. Culture finally played an overwhelming role in their adaptive success, making them menacing challengers for control of the entire biosphere. Most of the human odyssey is summed up in the following statement: more and more cultural adaptation, less and less biological adaptation, but some, nonetheless.

The notion of palaeoanthropology refers to the need to deal with both human palaeontology and palaeolithic archaeology in order to understand the extraordinary singularity of human evolution. This is the scope of *The Human Career*, but, surprisingly, of very few comparable publications. In its initial version, this book by Richard Klein was intended mainly for students. But after a decade of spectacular progress and controversies in this field, the book's format and contents have been completely revised, and it has been transformed from a textbook to a source-book. Its second edition stands out in the landscape of scientific publishing.

The Human Career describes one of the most spectacular changes to have occurred in our understanding of human evolution. The once-popular fresco showing a single file of marching hominids becoming ever more vertical, tall and hairless now appears to be a fiction. Humankind did not simply pass through successive stages, eventually leading

to the emergence of anatomically and behaviourally modern humans. For most of the past four million years, several species of hominids coexisted, sometimes in limited geographical areas. The eventual peopling of the planet with a single, homogeneous species of hominid is shown to be exceptional on the geological timescale. The phenomenon is especially noticeable with the first hominids, the australopithecines, for which eight species are now listed in the indexes. In the fierce controversies over the origin and relationships of these pre-humans, Klein is an outsider, developing his own views in a moderate style without becoming trapped in the usual pitfalls of either ignoring or trying to ridicule contradictory arguments.

A first 'out of Africa' movement followed the rise of a new model of hominid some 1.8 million years ago. *Homo ergaster*, a genuine human, was a tall, sweaty hunter adapted to open landscapes. This first movement resulted in the colonization of most of the warm and temperate zones of the Old World. *Homo erectus* developed in the Far East and Neanderthals in Europe. Eventually, one species, *Homo sapiens*, expanded to replace or outnumber its predecessors. The mechanism of the emergence of modern humans has been one of the most hotly debated issues in palaeoanthropology. Genetic and palaeontological evidence seems to support the African origin of all living populations. Yet questions remain over the path to modern humankind. Did certain archaic populations participate genetically in the development of the same living groups? How many 'out of Africa' events really occurred? What were the environmental and behavioural conditions of modern human expansion?

According to Klein, an 'out of Africa 2' movement started 50,000 years ago, spreading modern anatomy and behaviour throughout the planet. After a long period of very slow technical evolution, innovation and versatility suddenly became the engines of a genuine cultural revolution, while symbolic thought started to express itself. However, anatomically modern humans are known to have existed 100,000 years ago and did not develop any of these behaviours. Klein has his own answer to this puzzle. He favours the view that a neurological change occurred, rather than that purely linguistic, sociological and technical developments precipitated the revolution. This theory is hardly testable on the fossil record, and is where Klein will probably find most of his critics.

If you could have only one book that deals with human evolution, this is definitely the one to choose. In any event, the author has skilfully included all the others in it. If you are an aficionado, you will not want to miss this reference book describing the state of the art in a rapidly evolving field. ■

Jean-Jacques Hublin is at EP 1781, CNRS, 44 rue de l'Amiral Mouchez, 75014 Paris, France.

Science in culture

Photo-grass photographs

British artists win the L'Oréal Art and Science of Color Prize

Martin Kemp

On 25 January, at the Musée Mollot in Paris, the British artists Heather Ackroyd and Dan Harvey were presented with the Art and Science of Color Prize, awarded by the L'Oréal Art and Science Foundation. The prize was established in

1997 by the Tokyo-based foundation, the brainchild of Tetsuzo Kawamoto, for Japanese artists and scientists. It is intended to promote a creative dialogue between art, science and colour. This is its first international year, involving a multinational jury, including myself.

The winners, Ackroyd and Harvey, use grass as a novel surface for producing photographic images.

Anyone who has seen the effect made by something lying on a grass lawn for a few days will have noticed the yellowish imprint when it is lifted away. The pale silhouette may be considered as a kind of photographic negative, fading from view when sunlight restores the yellowed grass to its former greenness. Using the photosensitive property of grass, Ackroyd and Harvey have collaborated with scientists at the Institute of Grassland and Environmental Research (IGER), at Aberystwyth in Wales, to perfect the production of fully legible, 'living' photographs on a germinating lawn.

Their positive photographic images are formed by the projection of a black-and-white negative onto the surface of the growing grass. Such is the sensitivity of each germinating blade of grass that it produces a chlorophyll concentration that corresponds directly to the quantity of light available to it. Leaves of varied colour, from a rich, dark green to a sickly pale yellow, combine to form tonal images of a subtly elusive kind. They strongly recall the soft beauty of the callotypes made by William Henry Fox-Talbot during the years immediately following the first announcement to the public of the rival French and British photographic processes in 1839.

The biological researchers with whom Ackroyd and Harvey worked at IGER during 1997–98, as a result of a Wellcome 'Sciart' award, are developing techniques for controlling the



Ackroyd and Harvey's *Mother and child*, created by grass photosynthesis.

enzyme that degrades chlorophyll as a leaf dies. They have devised ways of modulating the expression of the gene responsible for the enzyme that causes the senescence of the green leaves. The scientists' newly engineered 'staygreens', which resist yellowing senescence, possess obvious potential for increasing the longevity of the artists' evanescent photographs. On their own behalf, Ackroyd and Harvey have placed particular demands on their unconventional medium in order to achieve minute gradations in green tonality. Their interests have been served by IGER's use of advanced visual analysis techniques, in particular the high-resolution imaging technique of hyperspectral analysis.

Ackroyd and Harvey's beguiling results, which revive the strange visual magic of early photographic images from more than 150 years ago, are underpinned by today's most advanced plant genetics and optical precision. Their grass-works testify to the ceaseless wonder of living nature in all its responsive subtlety, and to the more creative potential of human intervention in nature's supersensitive systems. Not least, it is a relief to find genetic engineering featuring in a story that does not involve scares — whether well- or ill-founded. ■

Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

New approaches to old age

To truly understand ageing, we must look beyond the diseases of old age.

Leonard Hayflick

The study of biological ageing was once the provenance of charlatans and snake-oil merchants. Over the last third of the twentieth century, however, it moved into the mainstream of biology. This change was probably driven more by the politics of the explosion in the numbers of old people in developed countries than by any advance in our understanding of the biology of ageing.

The purpose of much biomedical research is to increase life expectancy, and the twentieth century saw huge successes in this venture. In the past hundred years, life expectancy at birth in developed countries has increased from about 48 to 76 — the same gain that occurred over the previous 1,900 years. The average life expectancy at birth in ancient Rome of 20 resulted from the enormous number of deaths in early childhood, not unlike the present demographics in developing countries. It is the control of infectious diseases of the young that explains the increase in life expectancy during the twentieth century. Older people have benefited less because the chronic diseases to which they are more predisposed, such as cancer and cardiovascular disease, have been less yielding to biomedical research.

But this progress has neither advanced nor resulted from our understanding of ageing. Everyone knows the difference between the diseases that occur in early life and childhood development itself. Yet failure to distinguish between the diseases of old age and the ageing process is widespread even in the scientific community. The virtual resolution of poliomyelitis, acute lymphocytic leukaemia, Wilms' tumours and iron-deficiency anaemia did not increase our knowledge of childhood development; similarly, the resolution of the leading causes of death in old age — cardiovascular disease, stroke and cancer — are unlikely to advance our knowledge of the ageing process.

Changes attributable to disease can be distinguished from age changes because, unlike any disease, ageing occurs in virtually every species and only after reproductive success. Animals removed from the wild and protected by humans will undergo changes with age even though their species may have not experienced ageing for thousands or millions of years.

Humans, and the pet and zoo animals that we choose to protect, are the only species in which large numbers experience ageing. In the absence of human impacts on their

ecological niches, ageing in numbers proportional to those seen in humans simply does not occur in wild animals. The expression of age changes is not essential for the survival of any species. Humans have survived, and sometimes flourished, with a life expectancy at birth of 20 or 30 for more than 99.9 per cent of our time on the planet. Prehistoric human remains have rarely, if ever, revealed ages greater than about 50.

Human ageing has revealed itself because modern humans have escaped many causes of death not only in early life but long after reproducing. In so doing, we have unmasked a process for which evolution never prepared us. One might conclude that ageing is an artefact of civilization.

One example of the consequences for science policy of the failure to distinguish

We have unmasked a process for which evolution never prepared us. One might conclude that ageing is an artefact of civilization.

research on age-associated diseases from research on the fundamental biology of ageing is that it is virtually impossible to raise funds for research on ageing, because in the minds of policy-makers and the public no one suffers or dies from it. The general belief is that we suffer and die from the diseases that occur during the ageing process. Yet these diseases occur because age changes increase our vulnerability to what is ultimately written on death certificates.

More than half of the budget of the US National Institute on Aging is spent on Alzheimer's disease. Yet its elimination will have only a trivial impact on life expectancy and will not advance our knowledge of the fundamental biology of ageing. Greater attention must be given to a question that is rarely posed, yet which is capable not only of advancing our fundamental knowledge of ageing but also our understanding of age-associated diseases: why are old cells more vulnerable to disease than young cells?

The resolution of all causes of death currently written on the death certificates of those older than 65 will result in an increase in life expectancy of only about 15 years. An increase in our knowledge of how age changes occur does not put a 15-year limit on what is possible.

Leonard Hayflick is in the Department of Anatomy, University of California, San Francisco, PO Box 89, The Sea Ranch, California 95497, USA.



Generation gap: no one confuses child development and disease, but this is not true for the end of life.

HULTON GETTY

Pluto story

The discovery of extraterrestrial life was the key event of the third millennium.

Robert Silverberg

The discovery in AD 2668 of life on Pluto brought about humanity's greatest re-evaluation of its place in the Universe since the time of Copernicus, more than a thousand years before. It was Nicolaus Copernicus (1473–1543) whose astronomical calculations overthrew the ancient Ptolemaic theory of the heliocentric Solar System and demonstrated that the Earth was not the centre of the Universe but actually moves in orbit around the Sun.

Copernicus's work undermined the primacy of the Biblical view of the Universe and helped to weaken the religious establishment's power over scientific thought in mediaeval Europe. But the failure to produce evidence of life on any world but ours, even after the beginning of the age of space exploration, gave continued strength to the belief in the uniqueness of Earth.

The twentieth-century discovery of organic compounds in meteorites originating on Mars suggested that the red planet might once have been capable of sustaining life, but subsequent exploration offered no confirmation of this. The discovery late in the same century of a global ocean beneath the frozen surface of Jupiter's moon Europa aroused speculation that it might contain primitive life-forms, but this, too, proved untrue. And the numerous reports of visits to Earth by intelligent extraterrestrial beings, commonplace since the mid-twentieth century, have so far proved to be nothing

more than manifestations of popular irrationality.

Thus, by the middle centuries of the present millennium, most of us once more were convinced that Earth was the only place in the Universe where the miracle of life had ever occurred. There was no revival of the old churchly view that there had been a special act of creation: instead, it was generally thought that a unique and wildly improbable random event had taken place here on Earth — the blind shuffling of free molecules into a biological structure capable of persisting and reproducing itself. But this alone was enough to generate a kind of pre-Copernican mystical belief in the specialness of life on Earth. Though some iconoclasts warned that such thinking could lead to excessive complacency and an ultimate decadence, the absence of countervailing evidence robbed their arguments of any real force. Further exploration of space therefore seemed pointless, and hardly any took place in the deplorable 200-year period that began about 2400.

Then came the so-called Second Renaissance of the twenty-seventh century, bringing with it great prosperity and a revival of scientific curiosity. The inner planets of the Solar System were revisited after an absence of four centuries, and then the first voyages to the outer ones were made, culminating in the Pluto expedition of 2668 and the stunning discovery of living creatures there. "Pluto bears life", was the astounding, unforgettable message from the voyagers, who described crab-like creatures visible by the

thousand in the cold, glinting light of Pluto's day, scattered, as motionless as stones, along the shore of a methane sea, with thick, smooth, waxy-textured, grey shells and a great many jointed legs. No sign of life could be observed in them, even when they were prodded. But a few days later the bleak Plutonian night arrived, bringing with it a drop to two Kelvin, and they began to crawl slowly about. Evidently a state of dormancy was their norm except at temperatures of a few degrees above absolute zero.

Dissection of one captured specimen indicated an interior made up of rows of narrow tubes composed of silicon and cobalt lattices. A fluid flowing through these structures was identifiable as helium-2, the strange, friction-free form of the element found only at the extremely low temperatures typical of Pluto's night. Helium-2 makes possible the phenomenon known as superconductivity: the indefinite persistence of electrical currents flowing through a resistance-free medium. The obvious conclusion was that the energizing principle of the Plutonian creatures was superconductivity: that they were a life-form that could exist only on Pluto and function only during the Plutonian night.

But were these in fact life-forms? In the aftermath of the discovery it was widely argued that the crab-like things were nothing more than machines — mere signal-processing devices designed to operate at supercold temperatures, left behind, perhaps, by explorers from some other part of the Galaxy. Further study, however, indicated that the creatures performed the metabolic functions characteristic of life. They could be observed feeding on methane and excreting organic compounds. Apparent instances of reproduction by budding were also observed.

Today we have no doubt that the Plutonian creatures meet our definitions of true living beings. The myth of Earth's uniqueness in the Universe has been destroyed forever, and we are all familiar with the social and philosophical consequences. But are the Plutonians truly native to the frozen world where they were discovered, or are they sentinels posted there by some superior species from another star, which will return to our part of the Galaxy some day? Three centuries later that question remains unanswered, and we can only watch and wait.

Robert Silverberg has been a science-fiction writer for the past 45 years. His most recent books are *The Alien Years* and *Lord Prestimion (Voyager)*.



JACEY

Nogo in nerve regeneration

J. L. Goldberg and B. A. Barres

A protein long thought to inhibit the growth of regenerating nerve axons has at last been identified and cloned. This discovery could be an important step towards promoting nerve regeneration after stroke or spinal-cord injury.

Most tissues in the body — including muscle, skin, liver and peripheral nerve — have a remarkable ability to repair themselves after injury. Strangely, however, the brain and spinal cord, which constitute the central nervous system (CNS), have little innate capacity for repair. In adults, most parts of the CNS cannot generate new neurons or regenerate new axons, the thin processes that relay electrical impulses between neurons. This lack of regenerative ability explains the devastating permanence of injuries to the CNS, as occurs in spinal-cord paralysis.

Why don't adult CNS axons regenerate? Nearly a hundred years ago, Santiago Ramón y Cajal observed¹ that severed tips of CNS axons initially try to grow but then abort and fail to regenerate. Then, about 20 years ago, David and Aguayo² showed that some adult axons can grow through a peripheral nerve graft. These observations led to the hypothesis that axons fail to regenerate because they are inhibited by the adult CNS environment — specifically by non-neuronal 'glial' cells in the CNS known as oligodendrocytes and astrocytes. This possibility was elegantly tested by Schwab and colleagues^{3,4}, who discovered that dorsal root ganglion neurons in culture extend their axons across glial cells from the peripheral nervous system — Schwann cells — but avoid oligodendrocytes and the fatty myelin sheath that these cells wrap around CNS axons to provide electrical insulation.

These experiments raised the hope that axon regeneration might be promoted if the inhibitory oligodendrocyte molecules could be identified. That was Schwab and colleagues' goal when they identified two strongly inhibitory myelin proteins with relative molecular masses (M_r) of about 35,000 and 250,000. The authors then developed a monoclonal antibody named IN-1, which recognizes these proteins and can neutralize the inhibition of axons by the oligodendrocytes and myelin in culture⁵. When they injected IN-1 into adult rats after spinal-cord injury, about 5% of severed axons regenerated across the injured tissue and the rats showed striking functional improvements⁶. For a decade, however, the identity of the molecules recognized by the IN-1 antibody has remained a mystery.

Now, elsewhere in this issue (pages 383–384 and 434–444), three groups^{7–9}

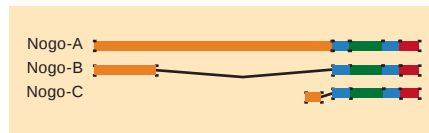


Figure 1 The three isoforms of Nogo. Predicted transmembrane domains are shown in blue, although it is not certain which portions of the molecule are extracellular. GrandPré *et al.*⁷ argue that only a 66-amino-acid domain (green) is located extracellularly, and show that it inhibits axons. Consistent with this, they find that human Nogo-C is also inhibitory. Prinjha *et al.*⁹, however, find that a construct encoding the amino-terminal domain (orange) of Nogo-A is inhibitory. Chen *et al.*⁸ show that an antiserum specific to a peptide sequence in this same domain neutralizes some of the myelin-associated inhibition¹¹. These observations indicate that there is an additional, extracellular inhibitory region in the amino-terminal domain. Perhaps Nogo-A has an unusual topology, or has two different topologies in the membrane.

report the identification in rats and humans of a hitherto unknown gene, *Nogo*, which encodes an inhibitory myelin protein. The breakthrough for all three groups came with the publication, by Schwab and colleagues¹⁰, of six partial-peptide sequences of a bovine inhibitory molecule that is recognized by IN-1. Chen *et al.*⁸ then used a rat expressed sequence tag to probe for similar sequences in a rat complementary DNA (cDNA) library; Prinjha *et al.*⁹ performed a similar manoeuvre to identify the human cDNA; and GrandPré *et al.*⁷ found a full-length human cDNA that had previously been deposited in the Genbank sequence database. Based on homology, the protein encoded by all of these sequences — Nogo — turns out to be a fourth member of the reticulon family⁷, which contains transmembrane proteins of unknown function. There are three alternative isoforms of Nogo, designated Nogo-A, -B and -C (Fig. 1).

Are all of the Nogo isoforms inhibitory? All three groups of authors agree that full-length Nogo-A protein has the strongest inhibitory activity in culture, and that it may be the M_r 250,000 protein recognized by the IN-1 antibody. As expected, Nogo-A is localized to CNS myelin, and is highly expressed by oligodendrocytes but not by Schwann cells. Nogo-B and -C are found in certain

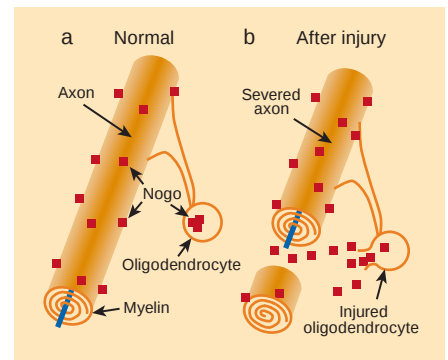


Figure 2 Inhibition of axon regeneration. a, Myelin-associated inhibitors, such as the newly discovered Nogo^{7–9}, are found on the inner and outer myelin leaflets and may normally prevent axonal sprouting in the uninjured CNS (also see Box 1, overleaf). b, After injury, damaged myelin and oligodendrocytes may prevent regeneration.

neurons and several non-neuronal tissues and, although their functions are not yet known, one of them may be the M_r 35,000 protein recognized by IN-1. There is some controversy, though, as to whether Nogo-B and -C are inhibitory, which Nogo-A domains are inhibitory, and what the configuration of Nogo-A is in the membrane (Fig. 1).

An initial surprise was that Nogo is largely found in a cellular organelle known as the endoplasmic reticulum. It contains a sequence that marks it to be retained there, where it cannot contact axons. So does any Nogo-A reach the surface of oligodendrocytes? At least two other myelin proteins also have retention sequences and yet make it to the myelin cell membrane. Both immunostaining^{7,11} and surface-biotinylation¹¹ experiments suggest that at least some Nogo-A reaches the cell surface of oligodendrocytes in culture. Moreover, two Nogo-A-specific antisera mimic the ability of IN-1 to neutralize the inhibitory effects of optic-nerve explants⁸ and oligodendrocytes on axon growth in culture¹¹. These findings indicate that at least some Nogo must be present on the cell surface (Box 1, overleaf).

But crucial questions remain. There is, as yet, no compelling evidence *in vivo* that Nogo-A prevents axons from regenerating, or that IN-1 antibodies enhance nerve regeneration by neutralizing Nogo. Although IN-1 neutralizes the inhibitory effects of Nogo-A in culture, this antibody binds to many other, unidentified spinal-cord proteins⁵.

Box 1 What do myelin inhibitors normally do?

The developing CNS is highly plastic, rapidly rewiring itself not only after injury but also normally, as it learns in response to environmental stimuli. But as the mammalian CNS ages, something mysterious happens, causing it to lose these abilities and stabilize its connections. Could myelin normally help to ensure the stability of neuronal circuitry by locking connections into place?

Consistent with this possibility, myelin develops *in vivo* at about the same time as the CNS loses plasticity, and it strongly inhibits axon growth. As well as increasing regenerative growth, the IN-1 antibody enhances the plasticity of axons in uninjured pathways, allowing them to grow into injured areas where they may help to restore function^{16,17}. Moreover, only older neurons — which may

have already reached and innervated their targets — are sensitive to inhibition by Nogo and by myelin-associated glycoprotein. Other known inhibitors, by contrast, repel younger axons as they grow towards their targets. With the discovery of the *Nogo* gene, we can now directly examine the normal functions of myelin in stabilizing the neuronal circuitry.

J.L.G. & B.A.B.

Numerous previously identified molecules could also, conceivably, inhibit regenerating axons. These include axon-guidance molecules such as semaphorins, ephrins and Slit¹², and other myelin molecules such as myelin-associated glycoprotein and certain proteoglycans. With the *Nogo* gene in hand, however, we can now look at its role in regeneration by determining whether specific anti-Nogo-A antibodies promote regeneration *in vivo*, as IN-1 does, and whether transgenic mice that lack the *Nogo* protein can regenerate their CNS axons better than normal mice.

If all the myelin inhibition *in vivo* could be overcome, would regeneration be assured? Davies *et al.*¹³ recently transplanted dorsal root ganglion neurons into myelinated regions of the CNS and found, surprisingly, that the neurons regenerated. So most of the myelin inhibitory activity might be released only after injury — a possibility that fits nicely with the largely intracellular location of *Nogo* (Fig. 2). Alternatively, Davies *et al.* suggest that reactive astrocytes found at injury sites, rather than myelin, inhibit axon regeneration.

But even if all glial inhibition could be overcome, there is another problem: neurons must be stimulated throughout life to survive and grow. In many regions of the CNS, when axons are cut, these stimulatory mechanisms are severely disturbed, leading to neuronal death and loss of regenerative ability¹⁴. So to promote regeneration we will probably need a multi-pronged approach in which inhibition is neutralized and growth is stimulated.

All that said, another startling experiment underscores the importance of myelin in preventing regeneration of axons in the CNS. David and colleagues¹⁵ found that if they immunized mice with myelin, there was at least ten times more nerve regeneration than previously seen. In half of these mice, roughly half the axons regenerated over long distances in the spinal cord, corresponding with considerable functional recovery. This remarkable finding hints at the enormous potential value of overcoming myelin inhibition.

Much work remains if we are to exploit these advances to help patients. But the discovery of the myelin inhibitory protein *Nogo* is a landmark step on a long road towards new treatments for patients suffering from many neurological conditions, including spinal-cord injury and stroke.

J. L. Goldberg and B. A. Barres are in the

Department of Neurobiology, Stanford University School of Medicine, Fairchild Science Building, Stanford, California 94305, USA.

e-mails: jlgoldbe@leland.stanford.edu
barres@stanford.edu

1. Ramón y Cajal, S. in *Degeneration and Regeneration of the Nervous System* (ed. May, R. M.) (Oxford Univ. Press, London, 1928).
2. David, S. & Aguayo, A. J. *Science* **214**, 931–933 (1981).
3. Schwab, M. E. & Thoenen, H. *J. Neurosci.* **5**, 2415–2423 (1985).
4. Schwab, M. E. & Caroni, P. *J. Neurosci.* **8**, 2381–2393 (1988).
5. Caroni, P. & Schwab, M. E. *Neuron* **1**, 85–96 (1988).
6. Bregman, B. S. *et al.* *Nature* **378**, 498–501 (1995).
7. GrandPré, T., Nakamura, F., Vartanian, T. & Strittmatter, S. M. *Nature* **403**, 439–444 (2000).
8. Chen, M. S. *et al.* *Nature* **403**, 434–439 (2000).
9. Prinjha, R. *et al.* *Nature* **403**, 383–384 (2000).
10. Spillmann, A. A., Bandtlow, C. E., Lottspeich, F., Keller, F. & Schwab, M. E. *J. Biol. Chem.* **273**, 19283–19293 (1998).
11. Huber, A. B., Chen, M. S., van der Haar, M. E. & Schwab, M. E. *Soc. Neurosci. Abstr.* **25**, 1263 (1999).
12. Chisholm, A. & Tessier-Lavigne, M. *Curr. Opin. Neurobiol.* **9**, 603–615 (1999).
13. Davies, S. J., Goucher, D. R., Doller, C. & Silver, J. *J. Neurosci.* **19**, 5810–5822 (1999).
14. Goldberg, J. L. & Barres, B. A. *Annu. Rev. Neurosci.* **23**, 579–612 (2000).
15. Huang, D. W., McKerracher, L., Braun, P. E. & David, S. *Neuron* **24**, 639–647 (1999).
16. Zagrebelsky, M. *et al.* *J. Neurosci.* **18**, 7912–7929 (1998).
17. Thallmair, M. *et al.* *Nature Neurosci.* **1**, 124–131 (1998).

Palaeoclimatology

Out of Africa

Frank Oldfield

A persistent and misleading myth is that climate variability has been minimal during the Holocene period, the present-day 'interglacial' that goes back 11,500 years. Striking evidence that this is not the case comes with the appearance of the paper by Verschuren *et al.* on page 410 of this issue¹.

The myth has arisen from the relatively invariant nature of the Holocene stable-isotope record (often used as an indirect measure of past air temperature) in ice cores from central Greenland, especially when compared with the record for the last glacial period². This is a polar-centric and exclusively temperature-orientated view of climate change. Had our perceptions been moulded instead by evidence from lower latitudes, with an emphasis on hydrological rather than temperature variability, the record from central Greenland would have appeared in a more realistic light. Several studies have confirmed the sequence of major changes in hydrological conditions in sub-Saharan Africa over the past 10,000 years³, but few have explored the record of hydrological variability on decadal timescales over the past 1,000 years.

Verschuren *et al.*¹ present just such a reconstruction with a high level of close chronological control. The reconstruction is based on detailed sediment studies from a small crater-lake basin in East Africa, with

sediment stratigraphy and fossil diatoms and midges being the mutually independent indicators of lake-level and salinity.

Simply as a piece of empirical research, the paper is of outstanding interest. The site lies in the kind of semi-arid region where human populations are most vulnerable to hydrological changes, especially persistent droughts. The results indicate that even during the past millennium, when the main pattern of climate conditions differed little from that prevailing before human-induced increases in greenhouse-gas concentrations in the atmosphere, there were at least three marked periods of extended drought: AD 1390–1420, 1560–1625 and 1760–1840.

In each case, the amplitude and duration greatly exceeded those of any droughts that have occurred during the short period for which instrumental records exist. The generally high degree of coherence between the different hydrological 'proxies' strengthens the authors' inferences. This demonstration of marked natural variability well outside the range captured by the instrumental record has a strong bearing on views about the future sustainability of human societies, because greenhouse-gas 'forcing' of climate will be superimposed upon and interact with such variability.

Verschuren *et al.* go further, and set their reconstructions of hydrological variability

alongside both the published record of past solar variability, inferred from changes in the production of ^{14}C , and the cultural record from the region (obtained from documentary records and oral history). Linking solar variability to climate change over the past millennium has gone in and out of fashion in recent years. However, there is a growing consensus that, despite the apparently slight shifts involved, the influence of changes in solar irradiance has probably been significant. These results¹ are therefore not inconsistent with current views, but they provide strong evidence for a link between solar and climate variability at the regional level and on decadal-to-century timescales.

The cultural record is in line with the history of drought evident in the sedimentary record. In particular, Verschuren *et al.* demonstrate an apparent correlation between periods of inferred high water levels and low salinity, and periods of 'prosperity' recorded in documents and oral history. This evidence comes in the wake of other studies from low latitudes that point to close links between human prosperity and changing climate⁴. Such indications should not encourage a reawakening of simple environmental determinism. But they should promote research into the relationships between past climate variability and human societies — not as a unidirectional line of causation, but as a set of highly interactive processes.

In this case¹, the apparent relationship between environmental variability and human prosperity persists through cultural changes spanning some 600 years. Understanding this kind of relationship calls for close collaboration between palaeo-scientists and anthropologists. There is every reason to encourage such collaboration, especially in areas that remain vulnerable to environmental stress.

Human societies suffer much more from declining or irregular water resources than from changes in temperature. Climate models often make a rather poor job of simulating even present-day precipitation or the balance between precipitation and evaporation. This is especially true at the regional scale, and at the level required to formulate models of the consequences of changes in water availability and of future sustainability in the face of such changes. By contrast, high-quality palaeo-data, where they are available with sufficient temporal resolution and dating control, are usually most convincing at the smaller spatial scale.

The findings of Verschuren *et al.*, then, are thought-provoking in both scientific and human terms. They illustrate the potential value that this form of research may have in sharpening our awareness of recent climate variability. More especially, they show how such research can help to provide more realistic projections of the effects of climate change

in regions where human societies remain hostage to an uncertain water supply. ■

Frank Oldfield is in the IGBP Past Global Changes (PAGES) International Project Office, Bärenplatz 2, CH-3011 Bern, Switzerland.

e-mail: frank.oldfield@pages.unibe.ch

1. Verschuren, D., Laird, K. R. & Cumming, B. F. *Nature* **403**, 410–414 (2000).
2. Dansgaard, W. *et al.* *Nature* **364**, 218–220 (1993).
3. Gasse, F. & van Campo, E. *Earth Planet. Sci. Lett.* **126**, 435–456 (1994).
4. Hoddell, D. A. *et al.* *Nature* **375**, 391–394 (1995).

Photosynthesis

Harvesting sunlight safely

Barbara Demmig-Adams and William W. Adams III

By harnessing the Sun's energy, photosynthesis provides the energy-rich and structural molecules (as well as oxygen) on which most life on this planet depends. Photosynthetic organisms, including green plants, have mastered the challenge of harvesting sunlight efficiently at one moment and then safely disposing of dangerous excess excitation energy the next. Without this protection through energy dissipation, photosynthesis as we know it could not occur in our relatively oxygen-rich atmosphere. The excess energy would instead be transferred from light-absorbing pigments to oxygen, resulting in lethal photo-oxidations by the same process that bleaches the colour from paintings exposed

to sunlight. Yet despite its importance, the mechanism of photoprotective energy dissipation is poorly understood. Now, through an elegant molecular approach, Li and co-workers¹ demonstrate (on page 391 of this issue) that a particular protein within the light-harvesting complexes of the photosynthetic apparatus has a unique, energy-dissipating function.

Plants often absorb more solar energy than they can use for photosynthesis — at noon, for example, when they are exposed to full sunlight (Fig. 1). When excited states of chlorophyll ($^1\text{Chl}^*$) are not readily processed by photochemistry (Fig. 2, overleaf), they can convert to triplet excited states ($^3\text{Chl}^*$), which can transfer their energy to oxygen. The resulting excited form of oxygen (singlet oxygen; $^1\text{O}_2^*$) is highly reactive and causes photo-oxidative damage (such as the peroxidation of membrane lipids) and, ultimately, cell death.

Normally, when excess light is absorbed, an alternative, dissipating pathway is activated that safely returns $^1\text{Chl}^*$ to its ground state before it can convert to $^3\text{Chl}^*$. The excitation energy of excess $^1\text{Chl}^*$ is thereby dissipated as heat (D in Figs 1 and 2) directly within the light-collecting chlorophyll- and carotenoid-binding protein complexes. In extreme cases when the rate of photosynthesis decreases to a negligible level — under environmental stress, say — this pathway can become the main, or only, route for excitation energy (Fig. 1). In this highly dissipating state, the light-collecting photosynthetic apparatus is effectively protected and can be preserved throughout periods of extreme stress^{2,3}.

Li and co-workers¹ used the fact that the level of thermal energy dissipation can be assessed non-invasively *in vivo* from changes in chlorophyll fluorescence emission. A small fraction of $^1\text{Chl}^*$ returns to the ground state, while re-emitting its energy as light in the form of fluorescence (Fig. 2). This chlorophyll fluorescence emission decreases in a characteristic way when the rate of thermal energy dissipation increases, and it can be quantified as non-photochemical quenching (NPQ) of chlorophyll fluorescence.

Much attention has been given to the role of carotenoids, particularly the xanthophyll

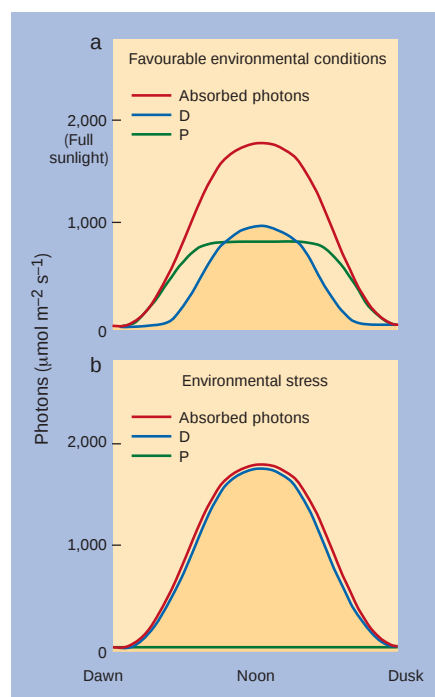


Figure 1 Allocation of absorbed sunlight to photosynthesis versus photoprotective dissipation. Use of absorbed photons in photochemistry (P; green) versus thermal dissipation of excess energy (D; red). a, On a clear day in sun-exposed leaves under favourable environmental conditions. b, In the presence of environmental stress, such as on a cold winter's day. A light intensity of $2,000 \mu\text{mol photons m}^{-2} \text{s}^{-1}$ corresponds to full sunlight.

zeaxanthin, in facilitating thermal energy dissipation⁴. For example, energy dissipation (NPQ) has been shown⁵ to be largely

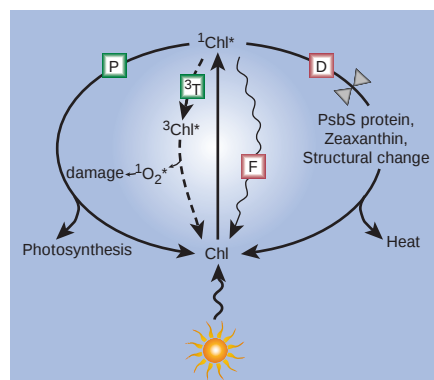


Figure 2 Fates of sunlight absorbed in the light-harvesting chlorophyll complexes. Chl, chlorophyll; $^1\text{Chl}^*$, excited singlet chlorophyll; $^3\text{Chl}^*$, excited triplet chlorophyll; P, photochemistry (green); D, safe dissipation of excess excitation energy as heat (red); F, fluorescence; ^3T , triplet pathway, leading to the formation of singlet oxygen ($^1\text{O}_2^*$) and photo-oxidative damage.

absent in zeaxanthin-deficient mutants, and it is impaired in mutants that lack lutein, a close isomer of zeaxanthin. But although the zeaxanthin-deficient mutants contained increased levels of vitamin E (α -tocopherol, which can detoxify singlet oxygen by accepting its energy and dissipating it as heat), they showed photo-oxidative damage in the form of lipid peroxidation⁶. When exposed to excess light under environmental stress, leaves of these mutants developed bleached, necrotic areas and died.

This central role of zeaxanthin (and possibly lutein) in the photoprotection of photosynthesis is paralleled by a function for these same two carotenoids in photoprotection of the human eye^{7,8}. In neither case, however, do we know the exact mechanism by which these carotenoids provide photoprotection. Indeed, we do not even know the location of thermal energy dissipation in the photosynthetic light-collecting system.

Li and colleagues' study¹ is a breakthrough because it shows that, among the many different pigment-binding proteins of the photosynthetic membrane, one specific protein is essential for thermal energy dissipation.

This protein, known as CP22 (chlorophyll-binding protein with a relative molecular mass of 22,000) or PsbS (the product of the photosystem 2 gene *psbS*), had not previously been implicated in this context. The authors isolated a mutant from the higher plant *Arabidopsis thaliana* that contained normal levels of zeaxanthin but was deficient in energy dissipation (NPQ). Li *et al.* then showed that this mutant lacks the PsbS protein. They used molecular and genetic markers to pinpoint the *psbS* gene and reintroduced a normal copy of this gene into the mutant plant, which regained the ability to dissipate energy thermally.

Photochemistry takes place in two photosystems, and numerous different light-harvesting protein complexes funnel energy into these photosystems. Over 30 distinct members of the light-harvesting complex family have been identified⁹, many of which are involved in efficient light harvesting. But the function of a subgroup—which includes the PsbS protein—has not yet been worked out. Li and co-workers¹ come to the important conclusion that photoprotective and light-collecting functions are not carried out

Citation analysis

Democracy on the rocks

Arguably the simplest way to review the state of a subject is not to commission an article by an expert reviewer, but just to make a statistical summary of what is going on in that field. Democracy does not have the same standing as a scientific methodology as it does as a principle for government, but would an objective summary of the subjects of published papers present a very different snapshot from an expert review?

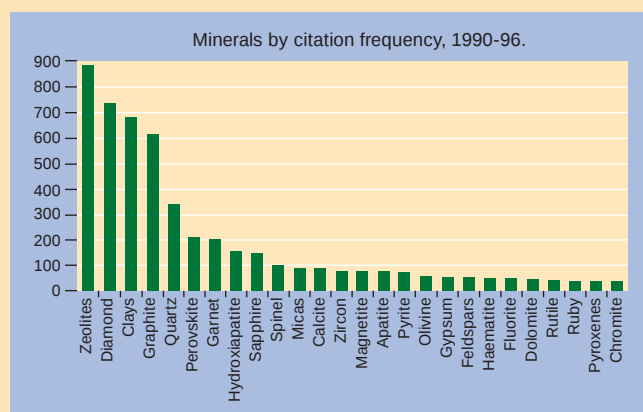
In an article entitled "The most popular rock stars, scientifically speaking" (*Annals of Improbable Research* (5, 20; 1999)), Vidal Barrón makes a light-hearted attempt to do just this for mineralogy. Barrón analysed the titles of papers in mineralogy published from 1990 to 1996, recording the frequency of occurrence of names of minerals or mineral groups to identify the "most popular". His results are summarized in a histogram, reproduced here.

Mineralogy emerges as a

subject whose remit swings rapidly from the slime to the sublime, with clays and diamond narrowly losing out to zeolites in the popularity stakes. This is an apt result because zeolites have filled both roles at different times—in the eighteenth century, fine zeolite specimens were among the most valuable minerals in the known world, but their prominence today is largely practical (for ion exchange and adsorption).

All five front-runners have important industrial applications, and clearly a high proportion of mineralogical research is being carried out in industry or at its behest. Number six, perovskite, is a curious star; this is a rare oxide that few mineralogists will have handled. The key to its popularity is that its densely packed structure is believed to be typical of mineral structures at high pressures in the mantle.

Overall, the list is dominated by industrial materials, with a smattering of gemstones and intellectually



challenging exotica. The main rock-forming minerals barely figure. An obvious anomaly is that minerals that are exploited in their own right feature strongly, unlike metallic ores, which usually occur in complex assemblages and so are unlikely to be named specifically in a paper's title.

What the list tells us is that mineralogy is a mature discipline, with the emphasis on applications rather than fundamental discoveries—probably not a balance that an expert review would have reached. The academic

emphasis is on minerals formed under extreme conditions in the inaccessible mantle. Considering that Earth scientists are proud of dealing with past events for which human activity is irrelevant and where hot silicates reign supreme, it is a little disconcerting to find that the top scorers are minerals rich in water or carbon—just as we are.

Bruce W. D. Yardley

Bruce W. D. Yardley is in the School of Earth Sciences, University of Leeds, Leeds LS2 9JT, UK.

e-mail: bruce@earth.leeds.ac.uk

by the same proteins. Their data also indicate that, during evolution, photoprotection arose before the appearance of highly efficient light-harvesting proteins.

Li and colleagues go on to show that the presence of PsbS is required for a conformational change in the photosynthetic membrane that engages thermal energy dissipation. At the same time, this conformational change also requires zeaxanthin and a pH gradient across the photosynthetic membrane. The authors discuss several models that could accommodate these findings. One attractive possibility is that protonation-dependent structural changes in PsbS facilitate downhill energy transfer from over-excited chlorophyll to zeaxanthin, followed by the rapid loss of this excitation energy from zeaxanthin in the form of heat, as is characteristic for carotenoids but not for chlorophylls.

What next? We now need to identify the exact location of the PsbS protein within the supramolecular structure of light-harvesting proteins. We also need to find out whether PsbS is, indeed, the site of thermal energy dissipation and the binding site for a specialized zeaxanthin molecule that facilitates most of the thermal dissipation. Alternatively, PsbS and the structural change in which it is involved may catalyse zeaxanthin-dependent energy dissipation in other

components of the light-collecting system.

Li *et al.* suggest that PsbS synthesis may increase in leaves exposed to higher levels of excess light on a daily basis. Before its function was known, increased levels of PsbS had also been observed³ in needles of Scots pine (*Pinus sylvestris* L.) in winter. These results show that we need to examine the involvement of PsbS in acclimation to the environment, where it probably modulates the maximal capacity for protective energy dissipation as needed. All of these questions should provide an exciting avenue for future research. ■

Barbara Demmig-Adams and William W. Adams III are in the Department of Environmental, Population, and Organismic Biology, Ramaley N122, University of Colorado, Boulder, Colorado 80309-0334, USA.

e-mail: barbara.demmig-adams@colorado.edu

1. Li, X.-P. *et al.* *Nature* **403**, 391–395 (2000).
2. Adams, W. W. III, Demmig-Adams, B., Verhoeven, A. S. & Barker, D. H. *Aus. J. Plant Physiol.* **22**, 261–276 (1995).
3. Ottander, C., Campbell, D. & Öquist, G. *Planta* **197**, 176–183 (1995).
4. Demmig-Adams, B. & Adams, W. W. III *Trends Plant Sci.* **1**, 21–26 (1996).
5. Niyogi, K. K. *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 333–359 (1999).
6. Havaux, M. & Niyogi, K. K. *Proc. Natl Acad. Sci. USA* **96**, 8762–8767 (1999).
7. Seddon, J. M. *et al.* *J. Am. Med. Assoc.* **272**, 1413–1420 (1994).
8. Chasan-Taber, L. *et al.* *Am. J. Clin. Nutr.* **70**, 509–516 (1999).
9. Jansson, S. *Trends Plant Sci.* **4**, 236–240 (1999).

Nanotechnology

One photon seen by one electron

Leo Kouwenhoven

Single photons can easily be detected by photomultipliers if their wavelength is much shorter than a micrometre. Photomultipliers contain photosensitive material inside a vacuum tube that produces a cascade of electrons when a photon strikes. This transformation from a single photon to many electrons works well in the visible and near-infrared range, but not for single photons of longer wavelengths.

Using a single-electron transistor as a detector, Komiyama *et al.*¹ report, on page 405 of this issue, measurements with unprecedented sensitivity for far-infrared (FIR) photons with wavelengths of several hundred micrometres. Such a detector is of great interest to spectroscopic research in the FIR region, which covers the vibrational spectra of molecules in liquids or gases, and there are many applications in astronomy, biochemistry and atmospheric research. The approach followed by Komiyama *et al.* marks a breakthrough because it provides single-photon sensitivity and simultaneously measures the wavelength of the photon.

There have been decades of research on quantum detection at submillimetre wave-

lengths² but, although detectors do exist in this regime, none have single-photon resolution³. A little thought about how individual quantum particles can be detected with a classical apparatus leads to the realization that this is far from trivial. What is actually happening inside a detector where at one end a particle, such as a photon, enters and at the other end a visible signal appears on a display? Somewhere inside, the small signal from the individual photon must have been immensely amplified to produce an electrical signal of around 10^{16} electrons per second (that is, milliamperes). This giant amplification should be accompanied by such a low noise level that unwanted errors do not occur — in other words, if there are no photons, there should be no signal. This becomes increasingly difficult for photons with smaller energy.

The energy range for submillimetre FIR radiation is somewhere between 1 and 100 millielectron volts (meV). As an example, a photon with a wavelength of 0.1 millimetres has a frequency of 3 terahertz and an energy of 12 meV. (For comparison, visible light is between 1.6 and 3.3 eV.) So, to obtain

wavelength selectivity, the detector should be sensitive in this energy range. To do this, Komiyama *et al.*¹ used cyclotron resonance between two Landau levels (Fig. 1). The Landau levels are narrow bands of allowed electronic states, which arise in a two-dimensional electron gas subjected to a large magnetic field. (In electron transport, the Landau levels lead to the quantized Hall effect.) Photon absorption can occur only when the photon energy equals the energy difference between the Landau levels. Under an applied magnetic field the energy separation changes by ~ 2 meV per tesla. So the magnetic field provides a knob to select a certain photon energy. Moreover, it transfers the signal of a single photon into an excitation of a single electron.

The next step is the detection of this photo-excited electron. Nowadays it is relatively simple to measure individual electrons by using single-electron transistors (SETs)⁴. These transistors are so small that the presence or absence of a charge from a single electron can switch the current on or off through the transistor. SETs can be made from metals or from semiconductors, in which case they are called quantum dots⁵. In fact, Cleland *et al.*⁶ have already reported single, FIR-photon detection using a metallic SET integrated with a silicon semiconductor structure. The

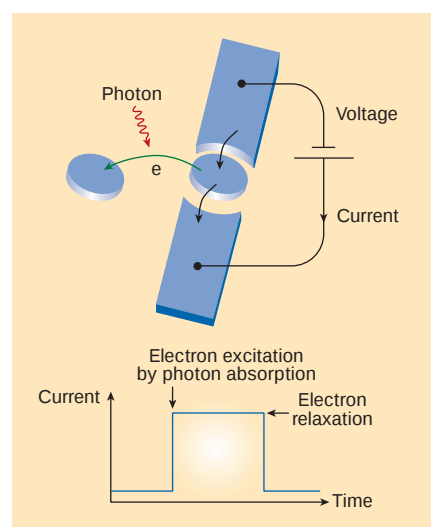


Figure 1 Single-photon detection by a single-electron switch, as demonstrated by Komiyama *et al.*¹. A quantum dot in the shape of a thin disk (typical radius of 100 nm) is connected through two electrodes to an external current and voltage circuit. In the presence of a high magnetic field, the electronic states of the quantum dot can be represented as two disks separated in space⁵. Initially the current is zero, owing to Coulomb blockade in the quantum dot. When a FIR photon (red arrow) excites an electron (green arrow), Coulomb blockade of the central disk is removed and current can flow (black arrows). The current switches to zero when the electron relaxes back to the central disk. The quantum dot is now ready to detect the next photon.

experiment by Komiyama *et al.* has much improved noise, but most importantly it is frequency, or equally wavelength, selective.

This new FIR detector will never be as easy to use as the table-top photomultiplier. It needs to be cooled down to very low temperatures, for two reasons. First, the FIR-photon energy is much lower than the thermal energy at room temperature (which is 25 meV at 300 K). This means that unwanted, background photons from room-temperature blackbody radiation will dominate the detector signal. To suppress the influence of blackbody radiation, the detector is cooled below 1 K. Second, the quantum dot used by Komiyama *et al.* operates properly only below about 1 K (ref. 5), so the detector will always need to be inside a helium cryogenic system. However, this is true for all sorts of FIR detectors, including those based on superconductivity³.

After any demonstration experiment, many hurdles to developing a useful product remain. In this experiment on single-photon

detection, not only the detector but also the radiation source was cooled to low temperatures¹. This is not practical for many spectroscopic applications, such as in vibrational spectroscopy of biological molecules. Fortunately, interstellar gases are naturally cooled to about 3 K so their radiation can be measured by using the detector outside the Earth's atmosphere. ■

Leo Kouwenhoven is in the Department of Applied Physics and the ERATO Mesoscopic Correlations Project, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands.
e-mail: leo@qt.tn.tudelft.nl

1. Komiyama, S., Astafiev, O., Antonov, V., Kutsuwa, T. & Hirai, H. *Nature* **403**, 405–407 (2000).
2. Tucker, J. R. & Feldman, M. J. *Rev. Mod. Phys.* **57**, 1055–1113 (1985).
3. Richards, P. L. *J. Appl. Phys.* **76**, 1–24 (1994).
4. Grabert, H. & Devoret, M. H. (eds) *Single Charge Tunneling* Series B Vol. 294 (Plenum, New York, 1991).
5. Kouwenhoven, L. P. *et al.* in *Electron Transport in Quantum Dots in Mesoscopic Electron Transport Series E* Vol. 435 (eds Sohn, L. L. *et al.*) 105–214 (Kluwer, Dordrecht, 1997).
6. Cleland, A. N., Esteve, D., Urbina, C. & Devoret, M. H. *Appl. Phys. Lett.* **61**, 2820–2822 (1992).

Global change

Rice, microbes and methane

Joshua Schimel

Methane is present at about only 1.8 parts per million in the atmosphere, but is a key player there—it is a greenhouse gas, it is central to atmospheric oxidation chemistry, and it is ultimately a source of stratospheric water vapour, which influences ozone depletion. Moreover, the concentration of methane is increasing rapidly. Hence the interest in the paper by Bodelier *et al.*, on page 421 of this issue¹, which deals with methane emissions from rice paddies.

Most of the methane in the Earth's atmosphere comes from biological processes, and rice paddies are one of the main sources. A large fraction of the methane produced in rice soils is consumed, however, being oxidized to carbon dioxide by methane-oxidizing bacteria (methanotrophs) in the soil, and so never makes it to the atmosphere. In upland soils, ammonium, which is formed naturally but is also a major constituent of nitrogen fertilizers, can inhibit methane oxidation and methanotroph growth. It has been a common assumption that this should occur in other ecosystems as well. So it comes as a surprise that Bodelier *et al.* find that, in rice-paddy soils, ammonium actually stimulates methane oxidation and methanotroph growth. This phenomenon may dominate the overall response of methane cycling to fertilization in rice-paddy ecosystems.

According to current estimates, rice agriculture will expand by up to 70% over the next 25 years to support the growing human

population². This will involve both increasing the area under cultivation and maximizing productivity by crop breeding and fertilizer management. Until now, it was thought that using nitrogen fertilizers on rice would increase trace-gas emissions. When nitrate-based fertilizers are used, much of the nitrate is denitrified, causing increased emissions of nitrous oxide, another potent greenhouse gas and ozone depleter. Ammonium fertilization also has the potential to increase methane emissions³—not only does it increase plant growth and carbon flow to methane-producing bacteria (Fig. 1a, overleaf), but it can also inhibit methane oxidation (Fig. 1c).

How does ammonium inhibit methane consumption? Several explanations have been proposed, the most solidly substantiated of which is competitive inhibition at the enzyme level^{4,5}. This occurs because, at the molecular scale, methane and ammonium are similar in size and structure. As a result, the enzyme that oxidizes methane (methane monooxygenase) can bind to ammonium and react with it (Fig. 1c). Because the possibility of competitive inhibition is fundamental to the biochemistry of methane oxidation, it was generally thought that inhibition should occur in rice paddies as well as in upland systems. In fact, it may, and the work of Bodelier *et al.* does not rule it out.

However, the extent of competitive inhibition is proportional to the ratio between



100 YEARS AGO

Expressions of opinion from political leaders as to the value of scientific advice, and the need for scientific methods in Government Departments, are so rare, that some remarks which Lord Rosebery made upon this subject at Chatham on Tuesday come almost as a surprise. We have over and over again referred to the lack of interest in the progress of science, and the disinclination to take advantage of available applications, shown by official authorities concerned with national affairs. Only recently some of the scientific lessons taught by the [Boer] war have been pointed out in these columns... and some of the services which a committee of men of science could render to the Government if their advice were asked have been indicated. From the subjoined extract from Lord Rosebery's speech it will be seen that he is in accord with the methods advocated in these columns. If the war leads to an acknowledgement of the value of scientific opinion, the result will be one upon which the nation may be sincerely congratulated.

From *Nature* 25 January 1900.

50 YEARS AGO

In the United States, wildfowl have been suffering from epidemic diseases, and, though there is no reason to believe that any such epidemics are likely to occur in the British Isles, investigations are being made into causes of mortality among wildfowl in Great Britain. Comparatively few wild ducks are picked up dead; but should any carcasses be found it is possible that the birds may have died from some disease and that a post-mortem examination would be of value. Carcasses should be forwarded to the Pathologist, Prosectorium, Zoological Society of London... as quickly as possible after finding, as the longer the post-mortem examination is delayed the less likelihood there is of determining the cause of death. The birds should be carefully wrapped in paper and packed in a box in order to comply with an international post office agreement, which stipulates that there must be no leakage from pathological specimens. The name and address of the sender, the place, and the locality and date of finding should be enclosed in the parcel. Postage will be refunded. Birds affected by oil should *not* be forwarded.

From *Nature* 28 January 1950.

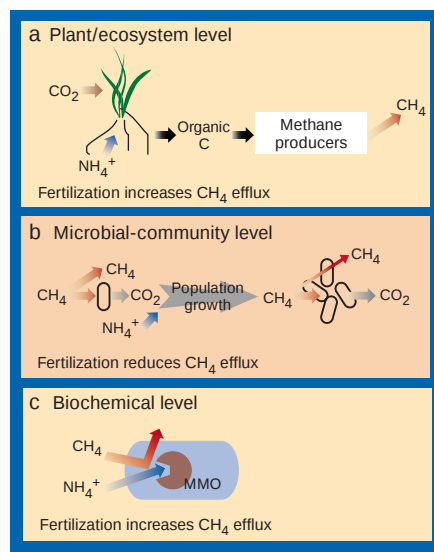


Figure 1 Effects of ammonium fertilizer on methane dynamics in a rice ecosystem. a, At the plant/ecosystem level, nitrogen increases plant growth and carbon supply to the methane producers, stimulating methane efflux from the soil. b, As Bodelier *et al.*¹ show, at the level of the microbial community, nitrogen stimulates the growth and activity of methane-oxidizing (methanotroph) bacteria, leading to reduced net efflux. c, At the biochemical level, ammonium inhibits methane consumption because of competition for methane monooxygenase, and so increases efflux. The balance between these processes may vary. But Bodelier *et al.*¹ find that effects at the level of the microbial community dominated the rice-ecosystem response to ammonium fertilization.

inhibitor and substrate. Methane concentrations are extremely high around a rice root, and plant nitrogen uptake may limit the exposure of bacteria to excess ammonium concentrations. These conditions should limit competitive inhibition. Equally importantly, because carbon (as methane) is readily available in rice soils, methanotrophs may become nitrogen-limited. So, at the level of the microbial community (Fig. 1b), addition of ammonium could stimulate methanotroph growth and activity to a much greater degree than it would inhibit them at the biochemical level (Fig. 1c). The net effect could be to reduce overall methane efflux from the ecosystem.

Bodelier *et al.*¹ show that, in a rice ecosystem, ammonium fertilization does indeed stimulate methanotroph growth and activity. Their research had three components: measurements of the rates of the reactions involved; studies using isotope tracers; and molecular analysis of the bacterial communities in the soil. The rate measurements showed a clear stimulation of methane consumption with fertilization. The tracer studies showed that, when ¹⁴C-methane was added, the ¹⁴C appeared in membrane lipids that are unique to methanotrophs. Finally,

DNA-based molecular analysis identified the species of methanotrophs present, and showed that the rice plant stimulated the growth of type I methanotrophs, the group that responded most favourably to nitrogen fertilization.

In a rice paddy, the interactions of nitrogen fertilizer and the methane cycle are complex, with different effects occurring at different levels of organization (Fig. 1). At the ecosystem and biochemical levels, fertilization would lead to increased methane emissions (Fig. 1a,c). As Bodelier *et al.* show, however, at the level of the microbial community, fertilization can stimulate methane consumption and reduce its efflux from the paddy (Fig. 1b). Which set of effects dominates the overall methane efflux may vary between systems, but responses at the level of the microbial community have generally not been considered. By incorporating these responses into their analysis, Bodelier *et al.* find that increasing rice production may not have such a severely detrimental effect on the atmosphere as has been assumed.

Further, their work highlights the importance of understanding ecological processes at the microbial community level. This is a level of organization that has been little studied. Probably less than 1% of bacterial

species have ever been isolated⁶, and many microbial processes have been studied in only a few ecosystem types. Only recently have researchers begun asking meaningful questions about how variations within physiological groups of microorganisms (methanotrophs, for instance) affect ecosystem-level processes such as trace-gas fluxes⁷. In large part, this has been because the molecular tools necessary for analysing microbial communities *in situ* have become available only in the past decade. Bodelier *et al.* have identified a gap in our understanding of microbial-community dynamics. As research continues, we are likely to find that some predictions about ecosystem behaviour fail because we are unaware of other such gaps.

Joshua Schimel is in the Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106, USA. e-mail: schimel@lifesci.ucsb.edu

1. Bodelier, P. L. E., Roslev, P., Henckel, T. & Frenzel, P. *Nature* **403**, 421–424 (1999).
2. Neue, H. U. *Soil Use Mgmt* **13**, 258–267 (1997).
3. Banik, A., Sen, M. & Sen, S. P. *Biol. Fertil. Soil* **21**, 319–322 (1996).
4. Dunfield, P. & Knowles, R. *Appl. Environ. Microbiol.* **61**, 3129–3135 (1995).
5. Gullledge, J. & Schimel, J. P. *Appl. Environ. Microbiol.* **64**, 4291–4298 (1998).
6. Amann, R. L., Ludwig, W. & Schliefer, K.-H. *Microbiol. Rev.* **59**, 143–169 (1995).
7. Schimel, J. P. & Gullledge, J. *Global Change Biol.* **4**, 745–758 (1998).

Fluid mechanics

Jets from a singular surface

Michael P. Brenner

All theories of continuous systems begin with the assumption that there is a separation of scales between the macroscopic variables needed by the theory and the microscopic motions that are ignored. So, the bulk motion of fluid in a teacup stirred by a spoon occurs on a scale that is orders of magnitude larger than the distance between the molecules. When this basic assumption is violated for an instant in time, the consequences can be important and dramatic. On page 401 of this issue, Zeff *et al.*¹ describe such a violation that occurs when a tub of viscous liquid is shaken up and down. When done just right, a relatively modest shaking can produce extremely thin and violent jets, which spout upwards at velocities of more than 50 metres per second.

The fact that violent jets can be produced in this fashion was first observed by Longuet-Higgins², who invented this experiment during a study of bubble cavitation. It has long been known that negative pressure in a liquid can create bubbles whose subsequent collapse can severely damage solid surfaces. The damage is believed to result from the impact of high-speed jets that form as the bubble collapses³. Longuet-Higgins's experiment

suggested that the creation of such jets from an oscillating free surface is a general phenomenon and that there may be a unique underlying mechanism. Over the years, there has been an emerging consensus^{4–6} in the theoretical community that jet formation in many different situations is coupled to a singular event on a free surface, such as that shown in Fig. 1.

Now, Zeff *et al.*¹ demonstrate experimentally that the violent eruption of a jet does indeed arise out of a singular distortion of the free surface: the shape of the liquid–air interface develops an extremely high curvature an instant before the jet is produced. The release of this curvature produces the jet. Zeff *et al.* show that this singular event is driven by the inertia and surface tension of the fluid, and ultimately is stopped by the retarding influence of viscosity. When the inertia of the fluid is above a critical threshold, the high surface curvature produces both a jet and an air bubble, which reduces the velocity of the jet.

The dynamics described by Zeff *et al.* are an example of a finite-time singularity, where the continuum description starts to break down. The formation of a singularity is accompanied by a divergence of a physical

quantity, which necessitates a change in the description of the system. A singularity is therefore a mixed blessing for a hydrodynamicist — on one hand it signals the breakdown of a cherished theory, but on the other the creation of a structure on an extremely small scale is a possibility to be exploited. The difference in scales between the singular structure and the rest of the problem means that the detailed properties of the singularity should be largely independent of the way the system is prepared. For this reason, understanding a singularity gives an opportunity to learn about a broader class of problems than was initially intended.

Historically, there are many examples where understanding singularities led to fundamental insights. Larson's theory of star formation, for example, is based on a singular solution to the equations of gravitationally forced gas dynamics⁷. Choptuik's study of singularity formation during gravitational collapse (that is, the creation of small black holes) argued against the 'cosmic censorship' hypothesis⁸, which states that singularities such as black holes are always concealed from observers. A particularly penetrating use of a singularity was shown by the British scientist G. I. Taylor, who calculated the energy of an atomic test explosion from a wartime United States government publicity movie⁹. Taylor did this without complete information of the explosion, because the blast wave has properties that are independent of the details of how the blast is created.

Singularities have one other special property: the shapes occurring before the singular event are exactly the same, after rescaling of the axes. This property is called self-similarity. Very simple arguments can often be given for the time dependence of the self-similarity, based solely on dimensional considerations. Zeff *et al.* find that the law governing the singularity on the free surface is identical to

that discovered by Keller and Miksis¹⁰ for a different problem (the recoiling of a fluid wedge on a solid surface) — this law applies whenever the most important forces are surface tension and inertia. Not only do Zeff *et al.* demonstrate that the Keller–Miksis law is obeyed in their experiments, but they also use it to construct a solution of the equations for fluid motion that is consistent with the observed shapes.

An obvious question posed by the history of this problem is whether the singularity observed has any relevance to jet production in cavitation damage. If a singularity obeying the Keller–Miksis law occurred during the collapse of a cavitation bubble, then the formation of violent jets could be a result of the release of high curvature on the bubble. The strongest argument for this appealing idea is that, because the solution describing the singularity is essentially independent of the details of the forcing, this same solution could in principle occur in other situations (such as cavitation) in which an air–liquid interface is strongly forced. On the other

hand, because the singularity seen by Zeff *et al.* occurs only at a carefully controlled value of the forcing, it seems unlikely that something so special will occur during the collapse of an arbitrary vapour bubble. The tension between these two points of view will need to be resolved with more research. ■

Michael P. Brenner is in the Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

e-mail: brenner@math.mit.edu

1. Zeff, B.W., Kleber, B., Fineberg, J. & Lathrop, D. P. *Nature* **403**, 401–404 (2000).
2. Longuet-Higgins, M. S. *J. Fluid Mech.* **127**, 103–121 (1983).
3. Plesset, M. & Prosperetti, A. *Annu. Rev. Fluid Mech.* **9**, 145–186 (1977).
4. Longuet-Higgins, M. S. & Oguz, H. *J. Fluid Mech.* **290**, 183–201 (1995).
5. Longuet-Higgins, M. S. & Oguz, H. *Phil. Trans. R. Soc. Lond. A* **355**, 625–639 (1997).
6. Cooker, M. J. & Peregrine, D. H. in *Proc. 6th Int. Workshop on Water Waves and Floating Bodies* 49–53 (1990).
7. Larson, R. *Mon. Not. Astron. Soc.* **145**, 271–295 (1969).
8. Choptuik, M. *Phys. Rev. Lett.* **70**, 9–12 (1993).
9. Taylor, G. I. *Proc. R. Soc. Lond. A* **201**, 159–186 (1950).
10. Keller, J. & Miksis, M. *Soc. Indust. Appl. Math. J.* **43**, 268–277 (1983).

Intelligence

Evolutionary psychology meets *g*

N. J. Mackintosh

Homo sapiens may not always seem to be wiser than other animals, but we are inclined to believe that we are more intelligent. Scepticism on this point is usually rebuffed by pointing to the hard evidence of the size of our brain and its threefold increase over the course of human evolution. But what is the nature of this superior intelligence given us by our large brains? And if intelligence has been of such importance to human evolution, why does it vary so much, at least as measured by IQ tests, among people today? Do IQ tests measure intelligence? If so, does the existence of the general factor *g*, extracted by factor analysis of any IQ-test battery, imply that there is a unitary trait of general intelligence? Or is the modular (Swiss army knife) view of the mind, beloved of evolutionary psychologists, a more accurate or fruitful metaphor?

These are large questions. If they were not all answered to everyone's satisfaction at a recent symposium*, there was a fair measure of consensus on some of them and much of interest to report on others.

The psychometricians were united in their support of *g*, or general intelligence. There are great differences in the surface features of different IQ tests or sub-tests: some ask simply for the meanings of words or items of general knowledge; others ask

for the solution of abstract, diagrammatic analogies or series-completion problems; and yet others require one to imagine what a complex three-dimensional object would look like viewed from a different perspective. Nonetheless, the fact is that performance on one kind of test correlates with performance on all others. This 'positive manifold' means that factor analysis will always yield a large general factor, which cannot, despite the views of some critics, be made to disappear by statistical legerdemain.

But what is the underlying cause of *g*? One popular suggestion (A. R. Jensen, Univ. California, Berkeley) is that there must be a biological basis, something like the efficiency of the brain as an information-processing device. There are (surely unsurprisingly) some correlations between IQ scores and some measures of the brain, for example its size. But although it has been pursued with much enthusiasm for 20 years or more, the search for correlations between IQ and any measures of brain activity, such as event-related potentials, has met with little success and yielded few replicable results. Moreover, the concept of brain 'efficiency' hardly has the theoretical precision needed to advance our understanding of the nature of intelligence.

A second approach has been to look for correlations between measures of general intelligence and performance on various

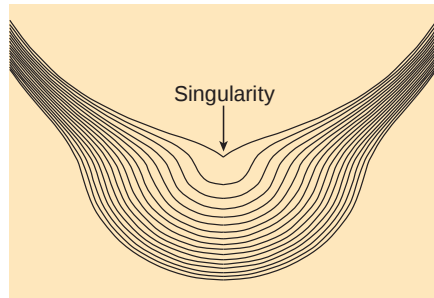


Figure 1 Formation of a singularity. Surface profiles from a model of a collapsing cavity, where the surface develops a singularity. The bottom profile corresponds to the earliest time, and the top profile to the latest. Zeff *et al.*¹ find a similar singularity occurring in a different situation, when a tub of viscous liquid is shaken up and down to produce violent jets. (Simulation by H. Oguz, Johns Hopkins University, and M. Longuet-Higgins, University of California at San Diego.)

*The Nature of Intelligence, Novartis Foundation, London, 30 November–2 December 1999.

elementary cognitive tasks (for instance, the reaction time needed to press the correct button when one or other of two lights comes on; or the time needed to decide whether two words are synonyms or antonyms, or whether a simple sentence is true or false) (I. Deary, Univ. Edinburgh). Here the message seems rather clear: performance on one of these tasks does not correlate with performance on others, but each independently correlates with *g*, usually at about 0.30. Together, however, they account for a sizeable portion of the total variance in *g* (D. K. Detterman, Case Western Reserve Univ.). The implication is that *g* depends on a whole series of relatively independent cognitive processes or operations.

This conclusion may well reconcile the apparent conflict between psychometricians, who study *g*, and evolutionary psychologists, who argue that the mind is modular. One way of understanding such modular theories is to define modules functionally: we, and our ancestors, have always had to solve a variety of problems, such as finding food, shelter or a mate, or recognizing other people, understanding their intentions and detecting whether they are lying. But rather than thinking of each of these tasks as being solved by an independent, encapsulated piece of neural machinery, hard-wired for this purpose, we should accept that we use multiple, overlapping sets of processes for each (R. M. Nesse, Univ. Michigan).

Is intelligence adaptive? And if it is, why does it still vary so much in the population when, according to Fisher's fundamental theorem, at equilibrium the additive genetic variance for any adaptive trait should be zero? At least in modern Western societies there has for some time been, if anything, a negative correlation between IQ and number of offspring. But there are many reasons why that is beside the point: among others, IQ may not measure the relevant aspects of adaptive intelligence, and to say that a trait has evolved because it is adaptive is to make an assertion about past adaptive significance, not current function. If we allow that modern hunter-gatherers provide the best contemporary evidence for earlier stages of human evolution, two lines of argument suggest that intelligence may have possessed adaptive significance (A. Whiten, Univ. St Andrews): first, it is clear that hunting depends on a variety of cognitive skills, not just on speed, strength or visual acuity; second, skilful hunters have more offspring than the less skilful — often with the less skilful hunters' partners.

This last observation has suggested the possibility that the evolution of intelligence has been subject to sexual selection (G. F. Miller, University College London), and that this might explain its persisting variation. Like the peacock's tail, intelligence, or its

overt manifestations, may be more a way of attracting the opposite sex rather than being of intrinsic adaptive value. Certainly, there is assortative mating for IQ (that is to say, a moderately high connection between husbands' and wives' IQ scores) and intelligence is rated as one of the most desirable characteristics in a mate in many cultures. But it remains to be shown that assortative mating for IQ is not largely a by-product of social stratification by IQ, and there are more mundane explanations (deleterious mutations,

say) for the persisting variations in IQ. According to Bailey's law (J. M. Bailey, Northwestern Univ.), the heritability of virtually all human traits is 0.50 ± 0.20 , and *g* is no exception. But the heritability of some other traits, such as schizophrenia, seems a great deal more problematic.

N. J. Mackintosh is in the Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK.
e-mail: n.mackintosh@psychol.cam.ac.uk

Earth science

Volcanic action at Axial Seamount

Earl Davis

Seafloor spreading is an intermittent process in which roughly 3 km³ of oceanic crustal rock is added to the Earth's surface each year along the global chain of mid-ocean ridges. But its intermittent nature, along with the remoteness and great depth of most spreading centres, makes the process difficult to study in 'real time'. An innovative approach is being taken by a group studying a submarine volcano called Axial Seamount, and they report their latest findings in a series of papers in *Geophysical Research Letters*¹⁻⁸.

Axial Seamount lies on the Juan de Fuca Ridge, a seafloor-spreading centre located about 300 km off the coast of Oregon, Washington and British Columbia. The investigators are using two approaches that are well suited to the study of active geological processes. The first is long-term continuous monitoring, with instruments deployed on the ocean floor and in the water column close to the site, as well as with several US Navy SOSUS coastal hydrophone arrays. These arrays were not designed for seismic monitoring, but are sufficiently broad-band and sensitive to locate offshore earthquakes that are too small to be detected by seismographs on land. The second approach involves 'event response', in which observations and experiments are carried out at the site as soon as possible after activity is indicated by the acoustic monitoring.

The name Axial Seamount is well chosen. The volcano lies at the intersection of the Juan de Fuca Ridge axis and a 'hotspot', a site of concentrated upwelling magma from the Earth's mantle. Although less intense, this hotspot is in many respects similar to those beneath Hawaii and Iceland, and it adds considerable fuel to the normal geothermal fire that feeds the rest of the Juan de Fuca Ridge. The seamount itself rises 1,000 m above the 'normal' ridge axis to the north and south (Fig. 1).

Since the beginning of real-time moni-

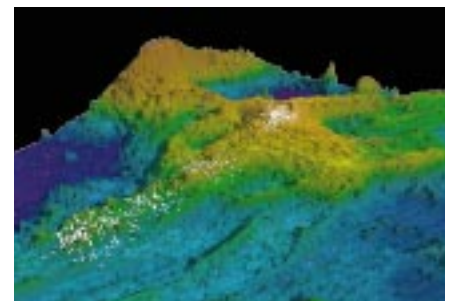


Figure 1 Bathymetry of Axial Seamount, the site of the January 1998 volcanic eruption along the Juan de Fuca Ridge described in the papers¹⁻⁸ discussed here. The view looks across a 100-km-long portion of the ridge, along a chain of seamounts striking northwest from Axial Seamount. Seismic activity associated with the eruption (epicentres shown as white dots¹) began with a cluster of earthquakes at the seamount's summit, then propagated down its southern rift zone (towards the left of the image). Water depths range from about 1,500 m (light brown) to 2,700 m (dark blue).

toring in 1991, two major episodes of seismic and volcanic activity have been detected by the SOSUS monitoring arrays. The latest began on 25 January 1998, and events associated with this eruption are described in the eight papers¹⁻⁸.

Action began with an earthquake 'swarm' beneath the floor of the summit crater, or caldera¹, with over 100 small-magnitude tremors occurring each hour (Fig. 2, overleaf). Only three were large enough to be evaluated using data from seismic monitoring stations on land. Within about three hours of the onset of seismic activity, an instrument bearing a pressure gauge identified the beginning of subsidence of the caldera floor² (Fig. 2). At the same time, an array of acoustic detectors set up across the rift zone running down the north flank of the volcano revealed lateral contraction of the seamount³, and sensors at the sea floor

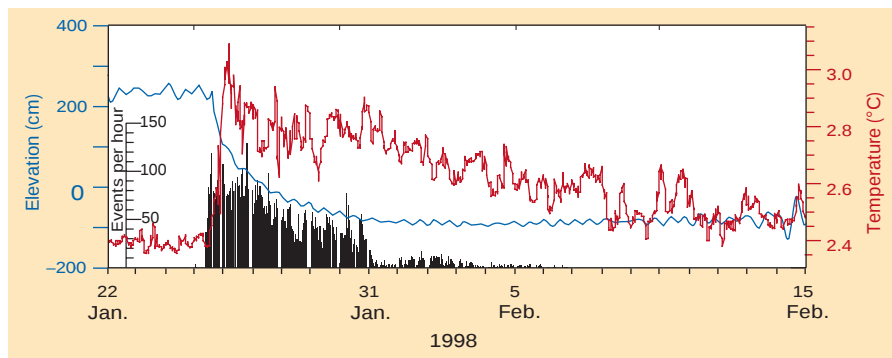


Figure 2 Sequence of eruption-related events at Axial Seamount from 22 January to 15 February 1998. Shown are the number of earthquakes identified per hour from SOSUS acoustic data (black)¹; subsidence of the sea floor of the seamount's caldera (arbitrary absolute scale) (blue)²; and temperatures measured 15 m above the sea floor (red)⁴.

and in the water column above the caldera detected a temperature anomaly^{2,4} (Fig. 2).

While the caldera remained active, seismic activity also began to migrate along the southern rift zone. Initially the rate was rapid, but it then stabilized to 20 km per day for about three days and ended roughly 70 km away (Fig. 2). Detectable activity continued for about ten days at the extremity of the rift zone, but at a much lower level¹. A low level of activity at the caldera was observed to continue for nearly two months with ocean-bottom hydrophones placed near the seamount summit⁸. The correlation between caldera subsidence and the migration of seismicity (Fig. 2) led to the conclusion that the eruption, fed along a sheet-like dike beneath the southern rift zone, had partially drained the magma chamber located about 3 km below the caldera floor.

Despite the possibility of winter storms, an immediate event-response cruise (followed by three more in the summer) was organized to take a closer look. The ship arrived 18 days after the initial detection of the earthquake swarm. Comparison of new swath-echo-sounding data and visual observations with previous information provided constraints on the thickness and extent of new volcanic rock erupted at the sea floor⁵. It is estimated from this and from the 3-m magnitude of the caldera subsidence that roughly $1\text{--}2 \times 10^8 \text{ m}^3$ of magma had been supplied from the summit magma chamber to seafloor flows and to the feeder dike beneath the southern rift zone during the eruption. Eruption of seafloor lavas extended 9 km down the rift, although this was only a fraction of the dike-propagation distance indicated by the seismic data.

Action at Axial Seamount was not confined to seismicity and volcanism. Initial signs of hydrothermal activity were seen in the monitoring data, and thermal profiling during the first response cruise revealed a large 'lens' of warm water southwest of the summit⁴, one with an anomalously low ³He/heat ratio⁶. This 'fingerprint' suggests that the warm lens had been produced

rapidly by heat exchange during the ten-day eruptive phase, and was subsequently swept some 20 km from its source by ocean currents. The increased rate of hydrothermal heat transfer was estimated to be as much as 200 gigawatts, more than two orders of magnitude greater than the typical heat output from the caldera during quiescent periods⁴.

One intriguing finding was the increase, relative to typical values in hydrothermal vent water at Axial Seamount, of methane concentrations in the plume created by the volcanic activity⁷. Although the source is unknown, one possibility is that micro-biologically 'mature' water in the crust, cool enough to support life, was suddenly heated by the volcanism and buoyantly ejected.

None of these observations runs counter to our general ideas about how volcanism and associated hydrothermal activity might operate. Nevertheless the data provide excellent quantitative information about the relationships between tectonics, volcanism and hydrothermal circulation at seafloor-spreading centres, and about the magnitude of one particular episode of activity at this site. They show that such events happen often enough for monitoring over a period of decades to bear results. And they make it clear that they are sufficiently short-lived to make continuous observatory monitoring essential. How typical this particular episode might be will emerge from continuing studies, and with expanded undersea technologies it will be possible to catch more of the action the next time it happens. ■

Earl Davis is at the Pacific Geoscience Centre, Geological Survey of Canada, Box 6000, Sidney, British Columbia V8L 4B2, Canada.
e-mail: davis@pgc.nrcan.gc.ca

1. Dziak, R. P. & Fox, C. G. *Geophys. Res. Lett.* **26**, 3429–3432 (1999).
2. Fox, C. G. *Geophys. Res. Lett.* **26**, 3437–3440 (1999).
3. Chadwick, W. W. Jr et al. *Geophys. Res. Lett.* **26**, 3441–3444 (1999).
4. Baker, E. T., Fox, C. G. & Cowen, J. P. *Geophys. Res. Lett.* **26**, 3445–3448 (1999).
5. Embley, R. W., Chadwick, W. W. Jr, Clague, D. & Stakes, D. *Geophys. Res. Lett.* **26**, 3425–3428 (1999).
6. Lupton, J. et al. *Geophys. Res. Lett.* **26**, 3449–3452 (1999).
7. McLaughlin-West, E. A. et al. *Geophys. Res. Lett.* **26**, 3453–3456 (1999).
8. Sohn, R. A., Crawford, W. C. & Webb, S. C. *Geophys. Res. Lett.* **26**, 3433–3436 (1999).

Daedalus

No more feelings

Last week Daedalus decided that consciousness, as a product of evolution, must be coded for in the genome of all conscious creatures. Anaesthetics abolish it selectively. 'Dissociative' anaesthetics, such as ketamine, even leave a degree of responsiveness. By identifying the gene or genes for consciousness, and working out the actions of the proteins they code for, DREADCO biochemists now plan to synthesize the ultimate dissociative anaesthetic. It will abolish awareness, but no other brain function. Like an alcoholic in a state of palimpsest, the user will seem entirely normal. But he will be a pure robot, reacting in all the usual ways but without feeling. Behind his fluent mannerisms and animated face, inside his skull, there will be nobody at home.

'Nothingness', as the new anaesthetic will be called, will bring compassion to modern farming. The ruthless brutalities of agri-business will still make its animal victims cry and cower in seeming misery. But these will be empty, robotic reactions, no longer denoting real suffering.

Human demand for Nothingness will also grow rapidly. It will be smuggled into prisons, releasing the inmates from their extended ordeal without brutal warders or interrogators suspecting anything. A prisoner on Nothingness will still eat, walk, talk, spit defiance or yell with pain without anyone suspecting that he is not suffering. People trapped in ghastly jobs or marriages, debilitating illnesses or grinding poverty, will also welcome the chance to erase their miseries while still fulfilling their obligations.

Nothingness will raise in an acute form the old philosophical problem of telling if anyone or anything is truly conscious, or is merely reacting without feeling. Alan Turing's famous test challenges a judge to distinguish the subject's responses from those of a computer simulation. In effect, the judge assesses the consciousness of man or machine by comparing them with an authentic sample of consciousness — his own. A robotic individual on Nothingness, with no internal standard of consciousness, could not do this. So Daedalus, cunningly, will judge the effectiveness of his product by a 'meta-Turing' test. A robotic, unconscious man will reveal the fact by being quite unable to judge a Turing test.

David Jones

The Further Inventions of Daedalus (Oxford University Press), 148 past Daedalus columns expanded and illustrated, is now on sale. Special Nature offer: m.curtis@nature.com

Energy for microbial life on Europa

A radiation-driven ecosystem on Jupiter's moon is not beyond the bounds of possibility.

The planet Jupiter's moon Europa may harbour a subsurface water ocean¹⁻³, but estimates of the available free energy have not been encouraging for supporting life^{4,5}. Here I show that disequilibrium chemistry in the ocean's ice cover, driven by charged particles accelerated in Jupiter's magnetosphere, should produce enough organic and oxidant molecules to fuel a substantial European biosphere. Microbial life could exist in concentrations detectable by surface landers able to filter meltwater from Europa's ice.

Europa's putative ocean lies beneath an ice layer too thick to permit photosynthesis¹. Hydrothermal vents may or may not exist. As terrestrial hydrothermal ecology depends largely on transported surface oxidants, it has been suggested that these are unlikely to exist on Europa². Some chemoautotrophic metabolism might occur, for example in methanogenesis from H₂ reacting with outgassed CO₂. But if carbon is outgassed on Europa as methane³, then most metabolic pathways operating on Earth may be denied to organisms inhabiting an ice-covered ocean on Europa⁵. I suggest instead that a radiation-driven ecosystem could exist, and calculate the lower limits on its biomass by reference to terrestrial analogues.

A 4.25- μm spectral feature may be due to CO₂ in Europa's ice layer⁶. Oxidants and organics should both be produced in H₂O/CO₂ ices through radiation chemistry⁷⁻⁹ and be retained in the ice matrix above normal sublimation temperatures by hydrogen bonding^{7,8}. Substantial radiation production extends down to about 1 mm (for a density, ρ , of about 1 g cm⁻³) (ref. 8), or to about 1 cm for $\rho = 0.1$ g cm⁻³. Sputtering removes 1 mm ($\rho = 1$ g cm⁻³) from the surface in roughly 5,000 yr (ref. 9). As the surface sputters away, the irradiated layer maintains a steady-state depth of 1 mm.

Charged-particle interactions with water should produce molecular oxygen, hydrogen peroxide and several other oxidants in comparable abundance⁷. Hydrogen peroxide has been detected¹⁰ on Europa at concentrations of 0.13% by number relative to H₂O. This concentration should hold throughout the top 1 mm, giving a column density of about 42 $\times 10^{18}$ H₂O₂ molecules cm⁻².

Cratering does not alter this model. The European crust appears to recycle on a 10⁷-yr timescale¹¹. Over 10⁷ yr, the turnover depth due to impacts is expected to be 1–10 cm (ref. 12), which will be sputtered away in only about 10⁵ yr.

Spectroscopy of Jupiter's moon Callisto suggests that CO₂ is present at ~ 0.5 wt%.

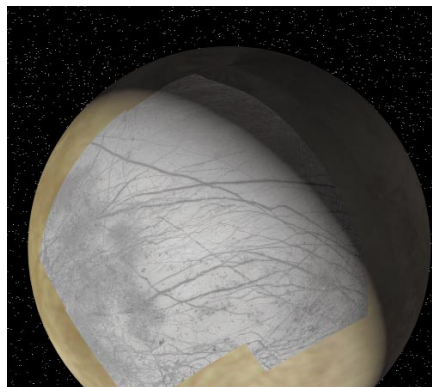


Figure 1 Jupiter's moon Europa is covered in ice, beneath which there may lie an ocean. Calculations suggest that, even without photosynthesis or hydrothermal vents, it could support life.

Europa's 4.25- μm spectral feature is $\sim 10\%$ of continuum globally, compared to a band strength of $\sim 25\%$ for Callisto⁶. Assuming a linear relationship between CO₂ concentrations and band strength for this unsaturated band suggests that CO₂ is present in European ice at ~ 0.2 wt% (T. McCord, personal communication). CO₂ in the upper 1 mm of Europa's surface will experience ~ 30 eV AMU⁻¹, integrating the flux⁷ of 32 $\times 10^{13}$ eV cm⁻² s⁻¹ over 5,000 yr.

Radiation should drive cycling among CO₂, CO and organics in the ice⁷⁻⁹. Organic groups may have been observed⁶. Laboratory irradiation of H₂O/CO₂ ices produces a dose-dependent CO/CO₂ ratio that is independent of the H₂O/CO₂ ratio¹³. Relative organic production may be calculated by comparing G values (molecules produced per 100 eV) for CO to those for organics⁷. At 30 eV AMU⁻¹, CO/CO₂ 0.3 (ref. 13). Formaldehyde and other simple organic species are produced as well. $G(\text{CO})/G(\text{HCHO}) \sim 4/9$ (ref. 7), so HCHO should be $\sim 32 \times 10^{15}$ by number in the upper 1 mm, corresponding to a column density, h , of about 82 $\times 10^{16}$ molecules cm⁻².

Organics and oxidants in Europa's crust are biologically relevant only if they reach the ocean. Crustal recycling mechanisms are unclear^{2,3}. Chaos regions may form by melt-through³. If resurfacing were primarily due to melt-through, about 12 $\times 10^{12}$ g HCHO, comparable amounts of other organics, and $\sim 72 \times 10^{13}$ g H₂O₂ would mix into the ocean every 10⁷ yr. The H₂O₂ in solution would decompose into H₂O and O₂ with a half-life of ~ 10 yr.

A putative microbial ecology on Europa could therefore be powered by the reaction $\text{HCHO} + \text{O}_2 \rightarrow \text{H}_2\text{O} + \text{CO}_2$. The terrestrial soil bacterium *Hyphomicrobium* can live on HCHO as its sole carbon source¹⁴. Terres-

trial methanotrophs obtain their energy by first oxidizing CH₄ to HCHO, then to HCOO⁻ and HCO₃⁻. Oxidation of HCHO in these organisms yields an energy of 4.7 eV per molecule¹⁴. The chemical energy available by melting the upper 1 mm of Europa's crust into the ocean is therefore 42×10^{17} eV cm⁻². If 10% of available chemical energy were used for biosynthesis, a typical terrestrial efficiency, w , for microbial biomass (dry weight) production is $w = 22 \times 10^{21}$ g eV⁻¹ (ref. 4). Biomass production on Europa would then be $42 \times 10^{17} \times 0.1 / 22 \times 10^{21} = 2 \times 10^{14}$ g cm⁻². Taking the dry mass of an aquatic cell¹⁵ to be 22 $\times 10^{14}$ g, this would give 42 $\times 10^{10}$ cells cm⁻².

If Europa's crust is recycled into the ocean over 10⁷ yr, net European cell synthesis is $dN/dt = 12 \times 10^{21}$ cells yr⁻¹, where N is the number of cells. The steady-state biomass is given by the product of dN/dt and the biological turnover time t . Adopting a turnover time appropriate for Earth's deep biosphere¹⁵, $t = 12 \times 10^3$ yr, $N = 12 \times 10^{24}$ cells, or $\sim 22 \times 10^{10}$ g. If biomass were not limited by energy, but instead by carbon available from HCHO alone, N would be roughly 10³ times smaller.

Radiation products might instead mix into Europa's ocean in sudden localized melting events analogous to Conamara chaos³. The mixing timescale can be estimated by comparing the energy flux from Europa's silicate/metal interior¹ to that required to melt the final 1 mm of ice: the required timescale is about 0.1 yr (probably an upper limit as melting events may involve larger-than-average local heat fluxes).

For an event of Conamara's area³, about 10⁴ km², a microbial mass $M = 2 \times 10^4$ km² $\times 82 \times 10^{10}$ g, or 42 $\times 10^{24}$ cells could bloom over ~ 0.1 yr, which is more than could be supported in steady state. This is $\sim 10^{14}$ of the Earth's oceanic prokaryotic population¹⁵. Microbial blooms could occur wherever Europa's surface communicated with its ocean — at chaos regions or along active cracks, for example. Any coloration or spectral features of such blooms after radiation processing could be explored by comparison with terrestrial analogues⁵.

European microbes would lie dormant and drift between melt-through or cracking events. There are at least ten chaos regions of $\sim 10^4$ km², or at least one of $\sim 10^5$ km² (ref. 3), and perhaps thousands of small melt-through events on Europa's surface³. If so, they probably occur frequently.

For an ocean about 100 km deep¹⁻³, average cell densities could be as high as 1 cell cm⁻³. If this water erupted to the

surface and froze, instruments being developed by NASA (P. Grunthaner, personal communication) could detect signs of life in 10 litres of melted and filtered surface ice. Ice from melt-through blooms may contain more cells, reducing melting requirements.

These calculations indicate that a particular radiation-driven ecosystem is plausible, quantifiable using our current knowledge of Europa: others are possible. Neither photosynthesis nor hydrothermal vents need be postulated. But only direct exploration will reveal whether life on Europa actually exists.

Christopher F. Chyba

SETI Institute, 2035 Landings Drive, Mountain View, California 94043, USA, and Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94035, USA
e-mail: chyba@seti.org

Physiology

A function for guttural pouches in the horse

Athletic animals must keep their brains cool during exertion because this organ can be damaged irreversibly by hyperthermia. But how horses do this has remained a mystery, as they don't appear to have thermoregulatory devices like those found in other animals. They do, however, have a unique anatomical arrangement of their internal carotid arteries, which supply blood to the brain: these are enveloped by a pair of air-filled guttural pouches. Here we show that horses use their guttural pouches to cool these important arteries during exercise, keeping the brain from overheating.

The guttural pouches are auditory-tube diverticula that contain about 300–500 ml of air. They are found in odd-toed hoofed animals, hyraxes¹, some bats and a South American forest mouse², and are susceptible to potentially life-threatening diseases from bacterial and fungal infections.

Their anatomical association with the upper respiratory tract suggests that the horse's guttural pouches might function during selective brain-cooling to maintain blood carried by the internal carotid arteries (ICA) at a temperature below the core temperature during hyperthermia. Inflowing arterial temperature has the greatest effect on the brain above 40 °C (ref. 3). Other selective brain-cooling devices include the carotid rete mirabile, which horses do not have, and a mechanism for direct cooling between the brain and cranial venous sinuses³.

Selective brain-cooling in horses is all but abolished by a tracheostomy⁴. The blood to the horse's brain is supplied mainly by the ICA, but also by the occipital and vertebral arteries⁵. The extracranial portion of the

1. Reynolds, R. T., Squyres, S. W., Colburn, D. S. & McKay, C. P. *Icarus* **56**, 246–254 (1983).
2. Pappalardo, R. T. et al. *J. Geophys. Res.* **104**, 24015–24055 (1999).
3. Greenberg, R. et al. *Icarus* **141**, 263–286 (1999).
4. Jakosky, B. M. & Shock, E. L. *J. Geophys. Res.* **103**, 19359–19364 (1998).
5. Gaidos, E. J. et al. *Science* **284**, 1631–1633 (1999).
6. McCord, T. B. et al. *J. Geophys. Res.* **103**, 8603–8626 (1998).
7. Delitsky, M. L. & Lane, A. L. *J. Geophys. Res.* **102**, 16385–16390 (1997).
8. Delitsky, M. L. & Lane, A. L. *J. Geophys. Res.* **103**, 31391–31403 (1998).
9. Johnson, R. E. in *Solar System Ices* (eds Schmitt, B. et al.) 303–334 (Kluwer, Dordrecht, 1998).
10. Carlson, R. W. et al. *Science* **283**, 2062–2064 (1999).
11. Zahnle, K., Dones, L. & Levison, H. F. *Icarus* **136**, 202–222 (1998).
12. Varnes, E. S. & Jakosky, B. M. in *Lunar Planet. Sci. abstr.* 1082 (Lunar and Planetary Institute, Houston, 1999) (CD-ROM).
13. Brucato, J. R., Palumbo, M. E. & Strazzulla, G. *Icarus* **125**, 135–144 (1997).
14. Madigan, M. T. et al. *Brock Biology of Microorganisms* (Prentice Hall, Upper Saddle River, New Jersey, 1997).
15. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).

ICA is enveloped by the thin (45–200 mm) mucous membrane of the guttural pouch's medial compartment (Fig. 1).

We measured the temperature of air from the guttural pouch as well as of blood from the ICA in four horses by using fine thermocouples (response times, 0.1 and 0.002–0.04 s) surgically implanted on the ICA at three sites corresponding to sites before, midway and beyond the guttural pouch, respectively (trifur, mid and f. lacerum) (Fig. 1a). At rest, trifur was significantly cooler than rectal temperature and mid (trifur, 0.565 ± 0.02 °C less than rectal; mid, 0.25 ± 0.03 °C more than trifur; results are combined mean ± standard error).

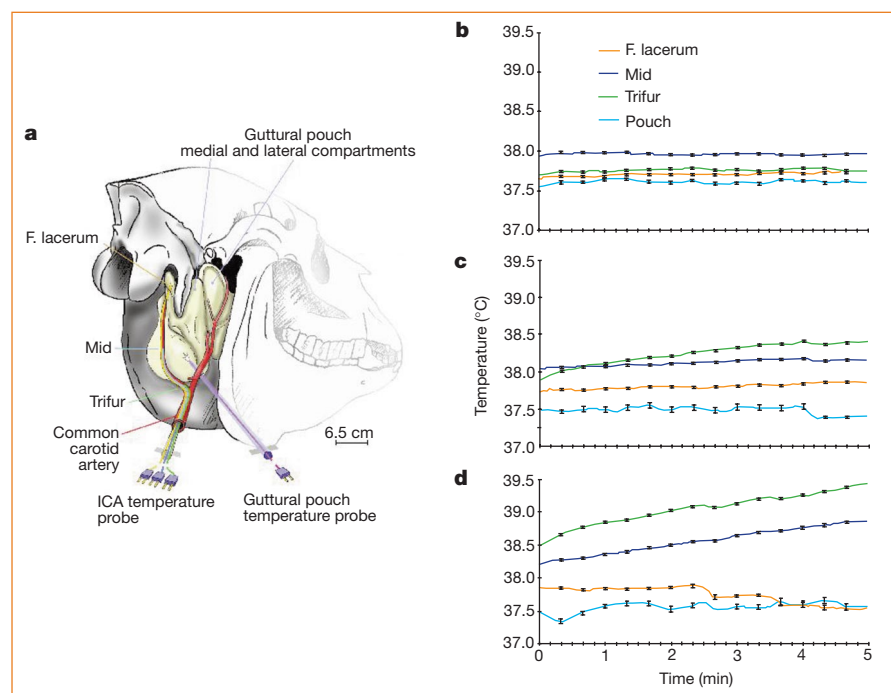


Figure 1 Guttural pouches act to cool the horses' brain. **a**, The arrangement of guttural pouches and internal carotid arteries (ICA) in the skull, and the position of the temperature probes used to measure blood temperature. **b–d**, Combined averaged temperatures (5 s.e.) for two trials in all four horses: **b**, at rest; **c**, trotting; **d**, cantering.

It has been reported that a horse's common carotid arteries are cooled⁶, presumably by heat loss from the pulmonary circulation and the proximity of the arteries to the trachea and skin. When the horse is resting, mid blood is warmed and f. lacerum blood cooled. Other surrounding tissues, notably the salivary glands and neck muscles, which are at core temperature, influence the warming of mid blood. Adjacent to f. lacerum is the extracranial ventral petrosal venous sinus, which is cooler than the resting core temperature⁴.

During trotting and cantering, the temperature at trifur rises, whereas that at mid and f. lacerum increases little by comparison (Fig. 1b–d). After prolonged cantering, the temperatures of guttural-pouch air and f. lacerum are similar. The rectal temperature differs from that of the ICA throughout, probably because of the lag phase between the horse's rectum and core⁴.

Guttural-pouch air temperature varied little with exercise (37.65 ± 0.03 °C at rest, 37.55 ± 0.05 °C trotting and 37.65 ± 0.08 °C cantering), as did rectal temperature (38.35 ± 0.01 °C at rest, 38.45 ± 0.04 °C trotting and 38.75 ± 0.1 °C cantering). Environmental air temperatures (and relative humidity) were 18.55 ± 3.5 °C (46 ± 5%) at rest, 18.35 ± 3.0 °C (45 ± 4%) trotting and 18.15 ± 2.7 °C (47 ± 5%) cantering. Exercise trials without the guttural-pouch temperature probe followed the same significant trends of cooling of the ICA.

Cool air in the guttural pouch would drive heat transfer from the passing ICA, and this is the only structure in this area

that is cooler than the blood. Experiments using radioactive microsphere tracers show that total blood flow in the brain is determined by the flow in the ICA⁵ and is not changed during heat exposure⁷ or exercise⁸.

Cantering for 15 minutes generates 15 kW of heat, of which 2.1 kW is dissipated by respiratory heat loss⁶. The maximum change recorded in the temperature of the ICA that could be attributed confidently to the guttural pouches is twice the mid change (2 °C). The total rate of heat transfer to each pouch is about 28 W (less than 0.4% of the total energy produced, assuming a blood flow of 3.5 ml s⁻¹ (ref. 5), blood density of 1,060 kg m⁻³ and blood specific-heat capacity of 3,700 J kg⁻¹ per °C (ref. 6)), despite its critical importance to selective brain cooling. This heat transfer from the ICA was minimal at rest but became more efficient with exercise.

It has also been proposed that selective brain cooling in the horse occurs through the intracranial cavernous venous sinuses (ICVS) cooling the ICA and brain base⁴. ICVS temperatures were cooler during hyperthermia, but the ICA temperature was not measured⁴. This mechanism is similar to that of the carotid rete mirabile, except that the horse's ICA have considerably less contact surface area and minimal arterial branching within the cooler venous sinus⁹. Cooling of the brain base is compromised by the thick dura mater between the horse's ICVS and brain. Given the horse's athleticism, and hence its thermoregulatory demands, an additional cooling mechanism is required: this is the role of the guttural pouches. Our results show that the ICA cools before entering the ICVS during exercise. Similar studies in the macaque monkey have shown that the ICA cools while passing the airways, before reaching the ICVS¹⁰.

Keith E. Baptiste*†, Jonathan M. Naylor*, Jeremy Bailey‡, Ernest M. Barber§, Klass Post*, Jim Thornhill¶

Departments of* Veterinary Internal Medicine, ‡Veterinary Anaesthesia, Radiology and Surgery, §Agricultural Engineering, and ¶Physiology, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5B4, Canada

†Present address: Danish Veterinary Laboratory, Bülowsvej 27, Copenhagen, Denmark
e-mail: keb@svs.dk

Neurobiology

Inhibitor of neurite outgrowth in humans

Axons are generally believed to be incapable of regeneration in the adult central nervous system^{1,2}. Inhibition results from physical barriers imposed by glial scars, a lack of neurotrophic factors, and growth-inhibitory molecules associated with myelin^{3–5}, the insulating axon sheath. These molecules include proteoglycans⁶, myelin-associated glycoprotein^{3,5} and, in bovine brain, two proteins called Nogo^{7–9}. We have used this bovine sequence to identify the human Nogo gene and have isolated complementary DNA clones encoding three different Nogo isoforms that are potent inhibitors of neurite outgrowth and which may help block the regeneration of the central nervous system in adults.

Using procedures described in Supple-

mentary Information, we generated a range of oligonucleotide primers to amplify and clone overlapping regions of the open reading frame (ORF) of Nogo cDNA. We identified three forms of cDNA clone (accession numbers AJ251383, AJ251384 and AJ251385 respectively) (Fig. 1a): the longest ORF of 1,192 amino acids (Nogo-A); an intermediate-length splice variant (Nogo-B) that lacks residues 186–1,004 in the putative extracellular domain; and the shortest splice variant (Nogo-C), previously described as rat vp20 (accession no. AF051335) and focen-s (accession no. AF132048), which also lacks residues 186–1,004 but which has a smaller, alternative amino-terminal domain (see Supplementary Information for full sequences).

The amino-terminal region shows no significant homology to any known protein, whereas the carboxy-terminal region shares significant homology with the neuroendocrine-specific proteins and other mem-

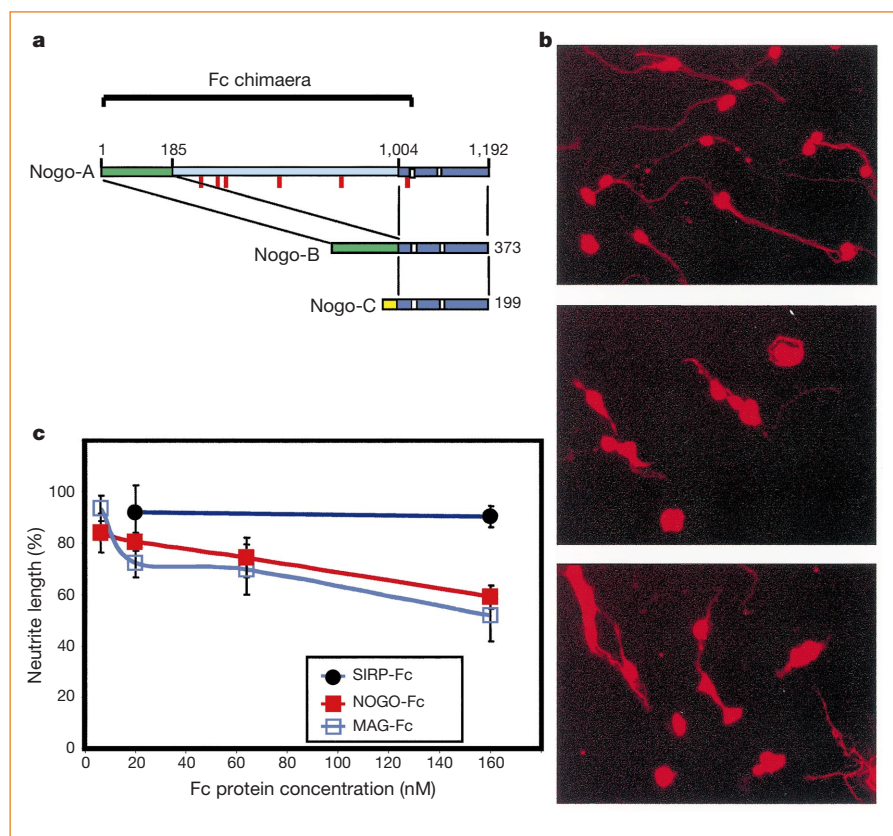


Figure 1 Predicted structure of Nogo isoforms and functional activity *in vitro*. **a**, Predicted structure of Nogo isoforms. Schematic diagram of the different variants of Nogo (A,B,C) that probably arise from alternative splicing. The relation between the three main cDNAs isolated is shown, with amino-acid residue numbering at key points (start, putative splice junction, stop). Predicted transmembrane regions are shown as white bars and the sequences corresponding to all six of the published Nogo peptide sequences are shown as red bars under the Nogo-A variant. The region of the Nogo-A polypeptide included in the Fc chimaera is indicated (residues 1–1,024). Pale blue, alternatively spliced exon unique to Nogo-A; dark blue, common C-terminal exon; green and yellow, alternatively spliced N-terminal exons. **b**, *In vitro* analysis of Nogo function. PB rat cerebellar granule neurons (CGCs) were dissociated and placed in culture on slides coated with poly-L-lysine and later supplemented with control Fc protein (SIRP-Fc) (top), or the inhibitory proteins MAG-Fc (middle) or Nogo-Fc (bottom). After growth for 48 hours, cells were fixed, permeabilized and stained with a neurofilament antibody. Micrographs of the treated cultures show the inhibitory effects of MAG-Fc and Nogo-Fc. **c**, Dose-response curve of CGCs treated as in **b** of Fc protein against neurite length. Results represent the mean neurite length per cell (5 s.d.) from two independent experiments, with more than 300 cells being measured in three separate wells for each treatment. MAG-Fc and Nogo-Fc each significantly inhibit neurite outgrowth at nanomolar concentrations.

- Brandt, J. F. *Bull. Acad. Imp. Sci. St Petersburg* **5**, 508–510 (1863).
- Hinchcliffe, R. & Pye, A. *J. Zool.* **157**, 277–288 (1969).
- Baker, M. A. *Annu. Rev. Physiol.* **44**, 85–96 (1982).
- McConaghy, F. F., Hales, J. R. S., Rose, R. J. & Hodgson, D. R. *J. Appl. Physiol.* **79**, 1849–1854 (1995).
- Orr, J. A., Wagerle, C., Kiorpes, A. L., Shirer, H. W. & Friesen, B. S. *Am. J. Physiol.* **244**, H142–H149 (1983).
- Hodgson, D. R. *et al.* *J. Appl. Physiol.* **74**, 1161–1170 (1993).
- McConaghy, F. F., Hodgson, D. R., Rose, R. J. & Hales R. S. *Equine Vet. J. (suppl.)* **22**, 42–46 (1996).
- Parks, C. M. & Manohar, M. *Am. J. Vet. Res.* **44**, 1861–1866 (1983).
- Ghoshal, N. G., Koch, T. & Popesko, P. in *The Venous Drainage of the Domestic Animals* (Saunders, Philadelphia, 1981).
- Day, M. *New Sci.* **161**, 7 (1999).

bers of the reticulon gene family, which are expressed in neurons mainly at intracellular sites^{10,11}. The carboxy-terminal tail contains a consensus sequence that may serve as an endoplasmic-reticulum retention signal.

From its sequence, Nogo may be a membrane-associated protein consisting of a putative large extracellular domain of 1,024 residues with seven predicted N-linked glycosylation sites (N-X-S/T), two or three transmembrane domains and a short carboxy-terminal region of 43 residues, but experimental mapping will be needed to determine its exact orientation in the membrane.

To investigate whether human Nogo inhibits neurite outgrowth like its bovine homologue⁹, we developed a soluble version of its largest isoform (relative molecular mass 220K). The extracellular region of this protein was prepared as a bivalent chimaeric protein using the CD33 signal sequence and human IgG Fc sequences. The purified protein was assayed *in vitro* for neurite-inhibitory activity. We allowed dissociated cerebellar granule neurons to adhere to substrate and then treated the cells with control Fc protein (SIRP-Fc), MAG-Fc or Nogo-Fc (Fig. 1b). We found that the control protein had no effect on neurite outgrowth, whereas myelin-associated glycoprotein (MAG) and Nogo were equally potent as dose-dependent inhibitors of nerve growth (Fig. 1c). The relative number of neurons present in each randomly selected field was similar for all the proteins tested, suggesting that the observed inhibition was not due to a failure of cell adhesion or indirect effects on cell survival.

Our results show that recombinant soluble Nogo produced as a bivalent Fc fragment is a potent neurite-outgrowth inhibitor. The availability of these active recombinant Nogo isoforms should enable their relative importance to be assessed in inhibiting neurite outgrowth at different anatomical loci and help in the identification of their receptor(s). Such definition is critical for the development of pharmacological treatments that will allow repair of lesions of the central nervous system.

Rabinder Prinjha, Stephen E. Moore, Mary Vinson, Sian Blake, Rachel Morrow, Gary Christie, David Michalovich*, David L. Simmons, Frank S. Walsh
Departments of Neuroscience Research and
*Bioinformatics, SmithKline Beecham
Pharmaceuticals, New Frontiers Science Park
(North), Third Avenue, Harlow,
Essex CM19 5AW, UK
e-mail: Frank_S_Walsh@sbphrd.com

1. Ramon y Cajal, S. *Degeneration and Regeneration of the Nervous System* (Oxford Univ. Press, London, 1928).
2. Schwab, M. E. & Bartholdi, D. *Physiol. Rev.* **76**, 319–370 (1996).
3. McKerracher, L. *et al. Neuron* **13**, 805–811 (1994).
4. Bregman, B. S. *et al. Nature* **378**, 498–501 (1995).
5. Cai, D., Shen, Y., De Bellard, M., Tang, S. & Filbin, M. T. *Neuron* **22**, 89–101 (1999).

6. Davies, S. J. A., Goucher, D. R., Doller, C. & Silver, J. *J. Neurosci.* **19**, 5810–5822 (1999).
7. Caroni, P. & Schwab, M. E. *Neuron* **1**, 85–96 (1988).
8. Rubin, B. P., Dusart, I. & Schwab, M. E. *J. Neurocytol.* **23**, 209–217 (1994).
9. Spillmann, A. A., Bandtlow, C. E., Lottspeich, F., Keller, F. & Schwab, M. E. *J. Biol. Chem.* **273**, 19283–19293 (1998).
10. van de Velde, H. J., Roebroek, A. J., Senden, N. H., Ramaekers, F. C. & Van de Ven, W. J. *J. Cell Sci.* **107**, 2403–2416 (1994).
11. Geisler, J. G., Stubbs, L. J., Wasserman, W. W. & Mucenski, M. L. *Mamm. Genome* **9**, 274–282 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as a paper copy from the London editorial office of Nature.

Materials

Creating the narrowest carbon nanotubes

The properties of carbon nanotubes¹ depend on their diameter and on the two integers (m, n) that describe their roll-up vector². The smallest nanotube reported previously had a diameter of 0.7 nm, the same as that of a C_{60} structure³, although nanotubes with a diameter of 0.4 nm have been predicted⁴. Here we report that simple improvements in the electric-arc technique can create a carbon nanotube with a diameter of 0.5 nm — the same as a C_{36} molecule⁵.

We used the same apparatus and conditions as described in ref. 6, except for the anode. This was prepared by boring a 3-mm-diameter hole in a 6-mm-diameter graphite rod. We filled these holes with a mixture of cobalt metal powder (5% at. metal) and the inner black core of a cathode deposit. This deposit was prepared

using a graphite anode, and consisted of carbon nanotubes and other amorphous carbonaceous materials. We used a high-resolution transmission electron microscope (HRTEM, JEM 200-cx, operated at 200 kV) to characterize the morphologies and microstructures of the products. The specimens for HRTEM analysis were prepared by dissolving the black deposit in ethanol. After ultrasonic treatment, a drop of the liquid was sprayed onto a holey carbon copper grid.

Figure 1 shows an HRTEM image of the carbon nanotubes found in our experiment. This shows a uniform intershell spacing of 0.34 nm. The open arrow in Fig. 1 indicates the innermost nanotube, with a diameter of about 0.5 nm. Although the entire tip is not clearly imaged, this innermost nanotube is closed at the end by a half circle, as indicated in Fig. 1. We suspect that it is closed by half a C_{36} cage. Theoretical results predict that a zigzag nanotube with this diameter might have an (m, n) value of (6,0) and a diameter of 0.47 nm or (7,0)(0.55 nm).

It has been suggested that carbon nanotubes are built up from atoms or atomic ions⁷. If this is true, anodes filled with graphite or carbon nanotubes should give the same results. Our results indicate that in the discharge the nanotubes could grow from carbon fragments. We propose that the carbon fragments within the arc have two forms before they begin to join up: curved (nanotube-filled anodes) and flat (graphite-filled anodes). Curved fragments will require less energy to form carbon nanotubes than flat ones. Use of a catalyst — cobalt, in our case — will result in curved fragments forming carbon nanotubes of the smallest diameter (0.5 nm in our experiment).

C_{60} has quite different properties from C_{36} : for example it is soluble in toluene, whereas C_{36} is not⁷, and the high curvature and increased strain energy of this smallest nanotube might lead to many unusual properties. These tubes are in the form of the innermost shell of multiwall nanotubes. An important question that remains is whether these small-diameter nanotubes could occur as single-wall nanotubes.

L. F. Sun, S. S. Xie, W. Liu, W. Y. Zhou, Z. Q. Liu, D. S. Tang, G. Wang, L. X. Qian
Institute of Physics, Center for Condensed Matter
Physics, Chinese Academy of Sciences,
Beijing 100080, China
e-mail: ssxie@aphy.iphy.ac.cn

1. Iijima, S. *Nature* **354**, 56–58 (1991).
2. Saito, R., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. *Appl. Phys. Lett.* **60**, 2204–2206 (1992).
3. Ajayan, P. M. & Iijima, S. *Nature* **358**, 23 (1992).
4. Sawada, S. & Hamada, N. *Solid State Commun.* **83**, 917–919 (1992).
5. Piskoti, C., Yarger, J. & Zettl, A. *Nature* **393**, 771–774 (1998).
6. Chang, B. H. *et al. Sci. China A* **41**, 431–436 (1998).
7. Endo, M., Iijima, S. & Dresselhaus, M. S. *Carbon Nanotubes* (Elsevier, Oxford, 1996).

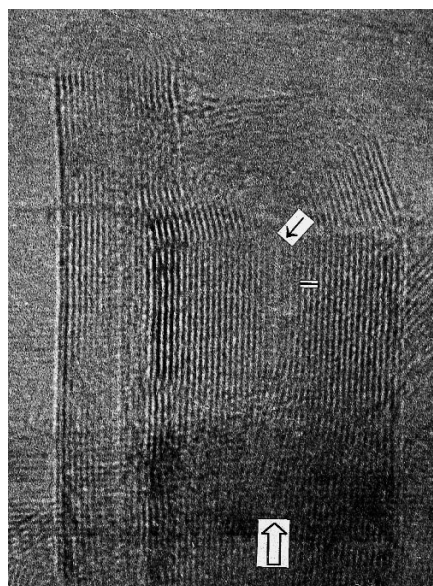


Figure 1 HRTEM image of two nanotubes with outer diameters of 6 nm and 16 nm, respectively. The hollow arrow points to the innermost tube in the right nanotube, which has a diameter of 0.5 nm. The other arrow points to the tip of this tube, which we believe to be closed with half a C_{36} cage. Scale bar, 1 nm.

Nodal signalling in vertebrate development

Alexander F. Schier* & Michael M. Shen†

* Developmental Genetics Program, Skirball Institute of Biomolecular Medicine, Department of Cell Biology, New York University School of Medicine, New York, New York 10016, USA

† Center for Advanced Biotechnology and Medicine and Department of Pediatrics, UMDNJ-Robert Wood Johnson Medical School, 679 Hoes Lane, Piscataway, New Jersey 10885, USA

Communication between cells during early embryogenesis establishes the basic organization of the vertebrate body plan. Recent work suggests that a signalling pathway centering on Nodal, a transforming growth factor β -related signal, is responsible for many of the events that configure the vertebrate embryo. The activity of Nodal signals is regulated extracellularly by EGF-CFC cofactors and antagonists of the Lefty and Cerberus families of proteins, allowing precise control of mesoderm and endoderm formation, the positioning of the anterior–posterior axis, neural patterning and left–right axis specification.

The vertebrate body plan is laid down during early embryogenesis, when the primary axes and tissue layers form. During the 20th century, classical studies have shown that embryogenesis proceeds through inductive interactions that are mediated by intercellular signalling factors, including members of the Wnt, fibroblast growth factor (FGF), Hedgehog and transforming growth factor- β (TGF β) families¹. Here we discuss recent progress suggesting that the TGF β -related factor Nodal may account for many of the roles attributed to TGF β activities during embryogenesis. Based on genetic evidence, we propose that (1) Nodal signals represent the *in vivo* inducers of mesoderm; (2) Nodal activity is modulated by cofactors of the EGF-CFC family as well as inhibitors of the Lefty and Cerberus families; and (3) Nodal signalling is mediated by Activin-like receptors and transcription factors of the Smad family. The regulation of Nodal signalling allows precise temporal and spatial control of mesoderm formation and is employed in several other critical patterning events in early development, notably in the determination of left–right (L–R) asymmetry.

TGF β -related activities induce mesoderm

The classic work of Nieuwkoop indicated that mesodermal cell types such as muscle and blood are induced by signals from neighbouring cells¹. Subsequent studies in *Xenopus* suggested that Activin and Vg1, two members of the TGF β family of signalling molecules, act as inducers of mesoderm *in vivo*¹. Activin was the first

TGF β molecule shown to have mesoderm-inducing activity using *in vitro* explant assays. Another strong candidate for a mesoderm-inducing factor has been the TGF β family member Vg1, on the basis of the localization of its messenger RNA to the vegetal hemisphere of the *Xenopus* embryo. This region has been implicated in inducing overlying cells to form mesoderm (Fig. 1). Moreover, processed and active Vg1 protein generated from bone morphogenetic protein (BMP)–Vg1 chimaeras can induce mesoderm in *Xenopus*. The identification of Vg1 and Activin suggested a relatively simple scenario in which maternally provided TGF β signals activate Ser/Thr kinase receptors and thus regulate mesoderm-specific genes. However, it has been difficult to show unequivocally that Activin or Vg1 are required for the formation of vertebrate mesoderm¹. For example, mouse mutants for both *Activin β A* and *Activin β B* form mesoderm and undergo normal gastrulation². Furthermore, blocking of Activin activity with the inhibitor Follistatin or with dominant-negative versions of Activin does not consistently block mesoderm formation in *Xenopus*^{3–6}. Although truncated activin receptors can inhibit mesoderm formation^{7,8}, these reagents are not fully specific for Activin as they can block BMP⁹, Vg1⁵ or Nodal¹⁰ signalling as well. In addition, processed Vg1 has not been detected *in vivo* and Vg1 mutant derivatives thought to block endogenous Vg1 activity cause only variable disruption of dorsal mesoderm formation¹¹. Similar experiments with the frog gene *Derriere*, which encodes a Vg1-related TGF β molecule, have indicated that it might be

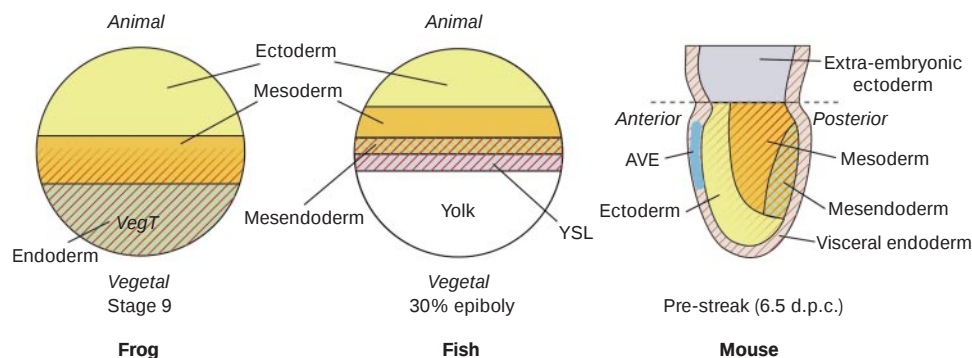


Figure 1 Schematic representation of Nodal signalling activities in frog, fish and mouse embryos before gastrulation. Expression patterns of *Nodal*-related genes (red hatching) are superimposed on simplified fate maps of embryos at the indicated stages. Mesoderm and endoderm are internalized through the primitive streak in mouse and the germ-ring in

fish. Classical embryological studies have shown that the vegetal region in frog (presumptive endoderm) and fish (yolk syncytial layer; YSL) are sources of mesoderm-inducing factors. *Nodal*-related genes are expressed in these regions. AVE, anterior visceral endoderm; d.p.c., days *post coitum*.

involved in specifying posterior regions, but is not required for general mesoderm induction⁶. It thus remains to be established whether members of the Activin or Vg1 families represent primary inducers of mesoderm *in vivo*.

Nodal is required for mesoderm formation

Genetic studies in mouse and zebrafish have shown that TGF β signals of the Nodal family are essential for mesoderm formation in vertebrates. Nodal was shown to be required for mesoderm formation by the isolation and cloning of a retroviral insertion mutation in the mouse *Nodal* gene^{12,13}, and the subsequent identification of a family of vertebrate *Nodal*-related TGF β genes^{14–20}. Mouse *Nodal* mutants lack a primitive streak (a thickening caused by ingression of mesoderm and endoderm progenitors) and most mesoderm, displaying only sporadic formation of some posterior mesoderm¹³ (Table 1). Similarly, double mutants for the zebrafish *Nodal*-related genes *cyclops* and *squint* lack head and trunk mesoderm and fail to form the germ-ring, a structure analogous to the mouse primitive streak²¹ (Table 1). *Nodal* genes are expressed in the vicinity of or overlapping with mesoderm progenitors (Fig. 1), including dorsal mesoderm precursors (also called the 'node' in mouse; hence the name *Nodal*). Moreover, mouse *Nodal*, zebrafish *Cyclops* and *Squint* and the frog homologues *Xnr1*, *Xnr2* and *Xnr4* can induce mesoderm formation in explant assays, similar to Activin and processed Vg1 (refs 15, 17, 19, 20, 22). Although additional signals (such as FGF and *Derriere*) may also be involved in (posterior) mesoderm induction, these genetic data establish that Nodal signals are central in the formation of vertebrate mesoderm.

Previous models have proposed that a maternal TGF β signal is involved in *Xenopus* mesoderm induction^{1,23}; however, *Nodal* genes are expressed zygotically during blastula stages (Fig. 1). This apparent contradiction appears to have been resolved by the recent identification of *Xenopus* *VegT* as a vegetally localized maternal T-box transcription factor. Depletion of *VegT* mRNA leads to a marked reduction in mesoderm and endoderm in *Xenopus*^{24,25}, and misexpression of *VegT* can induce *Nodal* gene expression^{26,27}. These results do not exclude a role for a maternally provided TGF β signal, but they suggest a revised model for mesoderm formation^{23,25} wherein a localized transcription factor(s) activates *Nodal* gene expression, which in turn leads to the induction of mesoderm. Determining whether a similar mechanism

controls mesoderm formation in fish and mouse will require the isolation of factors that activate expression of *Nodal* in these species.

Extracellular regulation of Nodal signalling

Recent studies have revealed an unexpectedly complex regulation of the Nodal signalling pathway, as shown by the identification of EGF-CFC proteins as Nodal cofactors and Lefty and Cerberus proteins as Nodal antagonists. It is now apparent that cells can become competent to respond to Nodal by expressing EGF-CFC proteins, and can attenuate Nodal activity by expressing inhibitors of the Lefty and Cerberus families (Fig. 2).

EGF-CFC factors. Members of the EGF-CFC family such as zebrafish *one-eyed pinhead* (*oep*) and mouse *Cripto* and *Cryptic* encode extracellular proteins that share an amino-terminal signal sequence, a divergent epidermal growth factor (EGF)-like motif, a second conserved cysteine-rich domain (the CFC motif) and a carboxy-terminal hydrophobic region^{28,29}. Although they were initially implicated in a receptor tyrosine kinase/ras pathway³⁰, genetic studies in zebrafish have shown that these extracellular proteins are essential for Nodal signalling³¹. Thus, embryos that lack both maternal and zygotic activity of *oep* display the same phenotype as double mutants for the *Nodal* factors *cyclops* and *squint*³¹ (Table 1). Strikingly, whereas overexpression of *Nodal* in wild-type embryos induces ectopic mesoderm, overexpressed *Nodal* in maternal-zygotic *oep* mutants has no phenotypic effect³¹. Indeed, similar to *Nodal* mutants, *Cripto* mutant mice do not form a primitive streak and embryonic mesoderm³² (Table 1). The cell-autonomous role^{31,33,34} and membrane association²⁹ of *Oep* indicate that EGF-CFC proteins might represent components of Nodal receptor complexes or otherwise facilitate the response to Nodal signals. In particular, Nodal signals might act upon cells only in the presence of EGF-CFC proteins. In turn, cells might be protected from responding to Nodal signals by lack of expression of EGF-CFC genes, thereby creating an additional level of regulation of these potent inducers.

Lefty. Members of the Lefty subfamily of TGF β molecules appear to act as antagonists of Nodal signalling. Lefty proteins are predicted to be structurally distinct from other TGF β molecules, lacking a long α -helix and a critical cysteine residue involved in stabilizing TGF β homodimers and heterodimers^{35–37}. It is thus possible that Lefty proteins act as monomers. Initially, Lefty proteins were thought to

Table 1 Selected genes in the Nodal pathway and their mutant phenotypes in mouse and zebrafish

Genes and genotypes	Species	Role and phenotypes	References
<i>Nodal</i>		TGF β -related factors; instructive signals	
<i>Nodal</i>	Mouse	Lack of primitive streak and most mesoderm; anterior neural truncations in chimaeras	12, 13
<i>cyclops</i>	Fish	Mild defects in prechordal mesoderm, lack of floor plate, cyclopia	17, 18, 20, 59
<i>squint</i>	Fish	Variable defects in axial mesoderm formation	17, 19, 21, 60
<i>cyclops; squint</i>	Fish	Lack of head and trunk mesoderm and endoderm, mispositioned anterior–posterior (A–P) axis	21
EGF-CFC		Extracellular cofactors for Nodal	
<i>Cripto</i>	Mouse	Lack of primitive streak, embryonic mesoderm and endoderm, mispositioned A–P axis	32
<i>Cryptic</i>	Mouse	Defective L–R axis formation (right isomerism and heterotaxia)	28, 53
zygotic <i>oep</i>	Fish	Lack of endoderm, prechordal mesoderm and floor plate, cyclopia	29, 33, 34, 57
maternal-zygotic <i>oep</i>	Fish	Lack of head and trunk mesoderm and endoderm, mispositioned A–P axis	31
partially rescued <i>oep</i>	Fish	Defective L–R axis formation (heterotaxia)	53
Lefty		TGF β -related factors; inhibitors of Nodal signalling	
<i>Lefty1</i>	Mouse	Defective L–R axis formation (left isomerism)	35, 36, 54
<i>Lefty2</i>	Mouse	Enlarged primitive streak, excess of mesoderm progenitors	10, 35, 36
<i>Lefty</i> overexpression	Fish	Lack of head and trunk mesoderm and endoderm, mispositioned A–P axis	10, 37, 38
Activin receptors		Putative cell-surface receptors for Nodal	
<i>ActRIIA</i>	Mouse	No early embryonic phenotype	61
<i>ActRIIB</i>	Mouse	Defective L–R axis formation (right isomerism)	51, 55
<i>ActRIIA</i> ^{−/−} ; <i>ActRIIB</i> ^{+/−}	Mouse	Variable defects in primitive streak formation, cyclopia, anterior neural truncations	51
<i>ActRIIA</i> ^{−/−} ; <i>ActRIIB</i> ^{−/−}	Mouse	Lack of primitive streak, mesoderm and endoderm	51
<i>ActRIIB</i>	Mouse	Lack of primitive streak, mesoderm and endoderm	50
Smad		Signal-transducing transcription factors	
<i>Smad2</i>	Mouse	Lack of primitive streak, embryonic mesoderm and endoderm	47–49
<i>Smad2</i> ^{+/−} ; <i>Nodal</i> ^{+/−}	Mouse	Variable defects in anterior midline, cyclopia, anterior neural truncations, L–R patterning	49
<i>Smad4</i>	Mouse	Lack of primitive streak, embryonic mesoderm and endoderm; anterior neural truncations in chimaeras	62, 63

be instructive signals or inhibitors of BMP signalling³⁶; however, recent studies in mouse and zebrafish indicate that these factors act as antagonists of Nodal, potentially by competing for binding to a common receptor (see below). In particular, the overexpression of mouse or zebrafish *Lefty* genes produces phenotypes in fish that are identical to those seen in *cyclops;squint* and maternal-zygotic *oep* mutants^{10,37,38} (Table 1). The phenotypic effects induced by *Lefty* can be countered by overexpression of *cyclops* or *squint*, suggesting an antagonistic relationship between Nodal signals and *Lefty* proteins^{10,38}. Similarly, *Lefty2* mutant mice have an enlarged primitive streak with more mesoderm progenitors (Table 1); this phenotype can be partially suppressed by decreasing Nodal activity, indicating that excess mesoderm formation may result from prolonged or hyperactivated Nodal signalling¹⁰. Interestingly, *Lefty2* expression temporally and spatially follows *Nodal* expression in mice, and studies in zebrafish indicate that Nodal signalling leads to the induction and maintenance of *Lefty* expression^{10,38}. Recent results indicate that a similar pathway may also operate in *Xenopus*³⁹. The negative feedback loop mediated by *Lefty* proteins effectively attenuates Nodal activity and renders it transient, allowing cells to respond to other signals after initial exposure to Nodal. This mechanism also blocks the positive autoregulation of *Nodal* expression¹⁰, possibly restricting the spatial range of Nodal activity. **Cerberus.** Overexpression studies in frogs have identified Cerberus as another molecule that can block Nodal signals^{40–42}. *Cerberus* encodes an extracellular protein that belongs to a subclass of the cysteine-knot superfamily, which also includes DAN and Gremlin⁴¹. Biochemical analysis and structural modelling have revealed that Cerberus-like proteins act as trifunctional growth-factor antagonists, which bind and block not only Nodal but also Wnt and BMP ligands^{41–43}. However, a truncated form of Cerberus (Cer-short) specifically blocks Nodal signals and, upon overexpression in *Xenopus*, induces phenotypes resembling *cyclops;squint* or maternal-zygotic *oep* mutants⁴². Although *Cerberus* expression does not

completely conform to Nodal signalling activity, it can be induced by Nodal signals in a subset of cells, thus potentially serving as a region-specific feedback inhibitor of Nodal signalling⁴². Nonetheless, an *in vivo* requirement for Cerberus-like proteins has not yet been established⁴⁴.

Transduction of Nodal signals

Although the *Nodal* signalling pathway has not yet been analysed biochemically, loss- and gain-of-function studies in mouse, frog and fish indicate that Nodal may utilize similar receptors and downstream effectors to those that have been identified for Activin (Fig. 2). Extensive studies have shown that Activin binds to the type II Ser/Thr kinase receptors ActRII (also known as ActRIIA) and ActRIIB, activating the type I receptor ActRIB, resulting in phosphorylation of Smad2. In turn, Smad2 forms a nuclear complex with Smad4 and members of the FAST family of forkhead domain transcription factors to regulate the expression of downstream target genes^{1,45,46}.

Several observations indicate that Nodal signals activate the same pathway as Activin or a highly related one. First, Nodal and Activin have similar activities as mesoderm inducers in explant assays and upon overexpression in whole embryos^{15,17,19,20}. Second, misexpression of activated versions of Activin receptors, Smad2 or FAST-1 produce effects similar to those of overexpression of Nodal signals and Activin^{1,46}. Third, maternal-zygotic *oep* mutants, which are defective in Nodal signalling, can be rescued by expression of activated forms of ActRIB and Smad2 (ref. 31). Strikingly, injection of *activin* mRNA can mimic this effect, indicating that forced activation of the Activin signalling pathway in *oep* mutants might recapitulate Nodal signalling *in vivo*, and that the combined activity of Nodal and EGF-CFCs may correspond to the activity of Activin alone³¹. Fourth, the extracellular domain of ActRIIB⁷ blocks the effects of overexpression of Squint, as well as the antagonistic effects of *Lefty*¹⁰. These results indicate that Nodal and *Lefty* may compete for interaction with a type II Activin receptor. Consistent with this model, *Lefty* also blocks Activin signalling³⁷. Finally, mouse *Smad2* and *ActRIB* mutants as well as *ActRIIA;ActRIIB* double mutants display gastrulation-defective phenotypes resembling those of mouse *Nodal* mutants^{47–51} (Table 1). As these mutant phenotypes are not identical, it is conceivable that Activin receptors and Smad2 also mediate additional signals during early embryogenesis, and that some aspects of Nodal signalling are mediated by additional downstream components. Taken together, however, the current genetic data are consistent with a pathway in which Nodal, in conjunction with EGF-CFC proteins, activates Activin receptors or Activin-like receptors and Smad2 or related Smad proteins, resulting in the transcription of downstream genes as well as *Lefty*, *Cerberus* and other potential feedback inhibitors, leading to the attenuation of Nodal signalling (Fig. 2).

Nodal signalling in left–right axis formation

The Nodal signalling pathway described above has been implicated in several other processes (Box 1), notably the specification of the L–R axis (Table 1; Fig. 2). In particular, members of the *Nodal* gene family in each vertebrate species (mouse *Nodal*; chick *cNR-1*; frog *Xnr1*; zebrafish *cyclops*) are specifically expressed in the left, but not right, lateral plate during somitogenesis; asymmetric expression is also observed around the node in chick and mouse⁵². The presence of Nodal signals in the lateral plate mesoderm has thus been proposed to determine 'left-sidedness', and absence of Nodal to determine 'right-sidedness'. In support of this model, ectopic expression of Nodal on the right side can perturb the asymmetry of internal organs⁵².

A requirement for Nodal signals in L–R asymmetry has been difficult to demonstrate, however, owing to the early embryonic lethality of *Nodal* mutants. Recent evidence for a direct role of Nodal signalling in L–R axis specification has been provided by analysis of

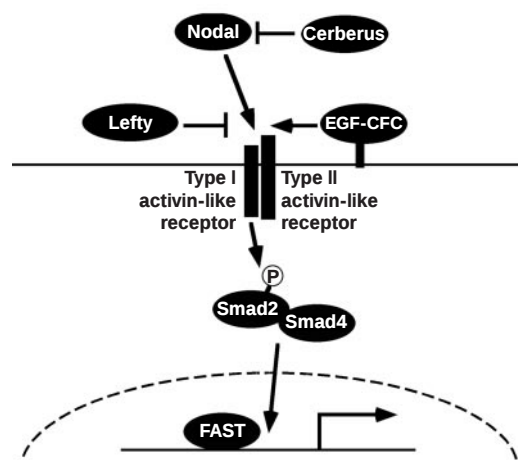


Figure 2 The putative Nodal signalling pathway. Loss- and gain-of-function studies indicate that Nodal signals may act with EGF-CFC cofactors to activate an Activin-like pathway. *Lefty* proteins act as feedback inhibitors during both mesoderm formation and L–R axis determination. Cerberus-related factors can bind directly to Nodal signals, but their role during development and their relationship with Nodal signalling is not clear. For instance, recent studies have identified a chick *Cerberus*-related gene, *Caronte*, which is expressed in the left paraxial and lateral-plate mesoderm^{43,64,65}. Although *Caronte* binds to Nodal protein, it unexpectedly activates *Nodal* gene expression in the lateral plate by inhibiting BMP activity. Furthermore, ectopic Nodal activity is unable to induce *Caronte* (at least in the trunk, though it is reported to do so in the head⁶⁵). Thus, in contrast to the proposed role of *Xenopus* Cerberus as an inhibitor of Nodal signalling during gastrulation, chick *Caronte* may not represent a downstream target or feedback inhibitor of Nodal signalling in L–R axis development.

the role of *EGF-CFC* genes in L–R axis formation in mouse and zebrafish. Mouse *Cryptic* and zebrafish *oep* are expressed in the lateral plate mesoderm, overlapping with *Nodal*, but are located symmetrically on both the left and right sides^{28,29}. Mouse *Cryptic* mutants display numerous L–R laterality defects including right-sided mirror-image symmetry of the lungs (right pulmonary iso-

merism), as well as loss of asymmetric gene expression in the lateral plate mesoderm⁵³ (Table 1). Moreover, zebrafish *oep* mutants that have been partially rescued by injection of *oep* mRNA show similar randomized laterality phenotypes and gene-expression patterns⁵³ (Table 1). These results are consistent with the idea that loss of Nodal signalling leads to 'right-sidedness' on both sides of the embryo.

Other genes implicated in the Nodal pathway are also required for L–R patterning. Mice lacking *Lefty1* often display left-sided mirror-image symmetry of the lungs (left pulmonary isomerism) and bilateral expression of genes that are normally expressed on the left⁵⁴. This phenotype is the opposite of that of *Cryptic* mutants, supporting the notion that *Lefty1* antagonizes EGF-CFC-dependent Nodal activity during L–R determination. *ActRIIB* mutants and *Nodal*^{+/-}; *Smad2*^{+/-} compound heterozygotes display L–R laterality defects that resemble those of *Cryptic* mutants^{49,55}. In summary, these findings indicate that during both mesoderm and L–R development, Nodal signalling is regulated by EGF-CFC and Lefty proteins and is mediated by ActRIIB and Smad2.

Future directions

We have discussed how a combination of instructive signals (Nodal), permissive cofactors (EGF-CFC) and feedback inhibitors (Lefty, Cerberus) regulates various aspects of vertebrate embryogenesis. The Nodal signalling pathway proposed above is based on genetic, not biochemical, evidence. First, therefore, it will be crucial to determine the molecular mechanisms by which Nodal signalling is regulated by EGF-CFC and Lefty factors and results in the activation of an Activin-like pathway. Second, *in vitro* studies on Activin have suggested that this TGF β signal can act as a graded long-range morphogen that determines distinct cell types at different concentrations and distances⁵⁶. Nodal signals can elicit concentration-dependent responses^{15,57} but, in contrast to Activin, Nodal might act at a short range⁵⁸. It is thus unclear whether Nodal signals pattern the embryo by acting as graded morphogens. Third, it remains a mystery how Nodal signalling controls development in different contexts. How is the identical pathway interpreted in distinct developmental processes, ranging from mesoderm induction to L–R patterning? Finally, overexpression of *Cripto* has been implicated in human carcinogenesis, particularly in the mammary gland³⁰. This raises the possibility that Nodal or related TGF β factors may be involved in human disease. More generally, the regulation of inductive signals by extracellular cofactors may be a general strategy employed during development and homeostasis. Extracellular cofactors such as EGF-CFC proteins may thus represent promising therapeutic targets for modulating the activity of signalling pathways. □

Box 1

Additional roles for the Nodal pathway

In addition to mesoderm formation and L–R axis specification, the Nodal signalling pathway has been implicated in several other processes in early vertebrate embryogenesis.

Endoderm formation. Mouse *Cripto* mutants lack definitive endoderm, similar to *cyclops*; *squint* double mutants and maternal-zygotic mutants for *oep*^{21,31,32}. The lack of a primitive streak in mouse *Nodal* mutants also suggests endoderm defects. In addition, recent work in fish and frog has implicated Nodal signalling as an early component in the pathway for endoderm formation^{26,27,66}.

Anterior–posterior patterning. Differential Nodal signalling might be involved in anterior–posterior (A–P) patterning in both mesoderm and ectoderm. Modulation of Nodal signalling activity in zebrafish indicates that specification of anterior axial mesoderm fates requires higher Nodal signalling than posterior fates⁵⁷. Increasing doses of zebrafish Lefty1 (Activin) progressively delete posterior neural fates, whereas overexpression of Nodal-related factors leads to posterior transformations⁵⁷. It is unclear whether high levels of Lefty1 specifically block Nodal or Activin signalling, but these results have led to the suggestion that Nodal-related signals and Activin specify fates along the A–P neural axis. Together, these results indicate that differential Nodal signalling in zebrafish may establish positional information along the animal–vegetal axis at blastula stages, which is then translated during gastrulation into distinct anterior–posterior mesodermal and ectodermal fates^{57,67}. In addition, a role for the Nodal signalling pathway in anterior neural induction has been suggested in mouse by analysis of *Nodal* chimaeras, *Smad4* chimaeras and *ActRIIA*^{-/-}; *ActRIIB*^{+/-} compound mutants, and in *Xenopus* by phenotypes induced upon overexpression of dominant-negative *Xnr2* derivatives^{51,62,68,69}. This analysis indicates that blocking Nodal signalling results in anterior truncations. However, this evidence for a role of Nodal in anterior neural induction is confounded by the observation that anterior neural tissue is present in *Cripto* mutant mice as well as maternal-zygotic *oep* and *cyclops*; *squint* mutant fish^{21,31,32}, and by the proposed role of the Nodal inhibitor Cerberus in head induction^{40,42}. A partial resolution of these issues might be that Nodal signalling is essential at an earlier step for visceral endoderm differentiation in the mouse, leading to upregulation of Cerberus and consequent head formation⁴².

Anterior–posterior axis positioning. Recent work in the mouse has suggested that the position of the A–P axis is established through a process of axial rotation involving morphogenetic movements of cells in the visceral endoderm and epiblast layers⁷⁰. An unexpected role for Nodal signalling in this process has been suggested by the finding that *Cripto* mutants display an A–P axis that is displaced by nearly 90° compared to that of wild-type embryos, consistent with failure of axis rotation³². Similar phenotypes have been observed in maternal-zygotic *oep* and *cyclops*; *squint* mutants, as well as in zebrafish embryos following overexpression of *Lefty*^{21,31,37}. These findings are consistent with the idea that an early Nodal signalling event is required to position the A–P axis in vertebrates.

Ventral midline formation. A role for Nodal signalling in formation of the ventral midline of the neural tube (including the floor plate) has been suggested by the phenotypes of *cyclops* and *oep* mutant fish^{18,20,33,34,59}. The cell-autonomy of *oep* in floor-plate progenitors suggests that Nodal is directly involved in the specification of floor-plate cells during gastrulation^{31,34}, in a step that does not depend on *sonic hedgehog*⁷¹. A role for Nodal in ventral forebrain formation is also supported by the cyclopia defects that are variably observed in *Nodal*^{+/-}; *Smad2*^{+/-} and *Nodal*^{+/-}; *ActRIIA*^{-/-} as well as *ActRIIA*^{-/-}; *ActRIIB*^{+/-} compound mutants^{49,51}.

1. Harland, R. & Gerhart, J. Formation and function of Spemann's organizer. *Annu. Rev. Cell Dev. Biol.* **13**, 611–667 (1997).
2. Matzuk, M. M. et al. Functional analysis of activins during mammalian development. *Nature* **374**, 354–356 (1995).
3. Hawley, S. H. et al. Disruption of BMP signals in embryonic *Xenopus* ectoderm leads to direct neural induction. *Genes Dev.* **9**, 2923–2935 (1995).
4. Kessler, D. S. & Melton, D. A. Induction of dorsal mesoderm by soluble, mature Vg1 protein. *Development* **121**, 2155–2164 (1995).
5. Schulte-Merker, S., Smith, J. C. & Dale, L. Effects of truncated activin and FGF receptors and of follistatin on the inducing activities of BVg1 and activin: does activin play a role in mesoderm induction? *EMBO J.* **13**, 3533–3541 (1994).
6. Sun, B. I. et al. *derriere*: a TGF- β family member required for posterior development in *Xenopus*. *Development* **126**, 1467–1482 (1999).
7. Dyson, S. & Gurdon, J. B. Activin signalling has a necessary function in *Xenopus* early development. *Curr. Biol.* **7**, 81–84 (1997).
8. Hemmati-Brivanlou, A. & Melton, D. A. A truncated activin receptor inhibits mesoderm induction and formation of axial structures in *Xenopus* embryos. *Nature* **359**, 609–614 (1992).
9. Hemmati-Brivanlou, A. & Thomsen, G. H. Ventral mesodermal patterning in *Xenopus* embryos: expression patterns and activities of BMP-2 and BMP-4. *Dev. Genet.* **17**, 78–89 (1995).
10. Meno, C. et al. Mouse Lefty2 and zebrafish activin are feedback inhibitors of Nodal signaling during vertebrate gastrulation. *Mol. Cell* **4**, 287–298 (1999).
11. Joseph, E. M. & Melton, D. A. Mutant Vg1 ligands disrupt endoderm and mesoderm formation in *Xenopus* embryos. *Development* **125**, 2677–2685 (1998).

12. Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L. & Kuehn, M. R. Nodal is a novel TGF- β -like gene expressed in the mouse node during gastrulation. *Nature* **361**, 543–547 (1993).
13. Conlon, F. L. *et al.* A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development* **120**, 1919–1928 (1994).
14. Smith, W. C., McKendry, R., Ribisi, S. J. & Harland, R. M. A nodal-related gene defines a physical and functional domain within the Spemann organizer. *Cell* **82**, 37–46 (1995).
15. Jones, C. M., Kuehn, M. R., Hogan, B. L., Smith, J. C. & Wright, C. V. Nodal-related signals induce axial mesoderm and dorsalize mesoderm during gastrulation. *Development* **121**, 3651–3662 (1995).
16. Levin, M., Johnson, R. L., Stern, C. D., Kuehn, M. & Tabin, C. A molecular pathway determining left–right asymmetry in chick embryogenesis. *Cell* **82**, 803–814 (1995).
17. Rebagliati, M. R., Toyama, R., Fricke, C., Haffter, P. & Dawid, I. B. Zebrafish nodal-related genes are implicated in axial patterning and establishing left–right asymmetry. *Dev. Biol.* **199**, 261–272 (1998).
18. Rebagliati, M. R., Toyama, R., Haffter, P. & Dawid, I. B. *cyclops* encodes a nodal-related factor involved in midline signaling. *Proc. Natl Acad. Sci. USA* **95**, 9932–9937 (1998).
19. Erter, C. E., Solnica-Krezel, L. & Wright, C. V. Zebrafish nodal-related 2 encodes an early mesendodermal inducer signaling from the extraembryonic yolk syncytial layer. *Dev. Biol.* **204**, 361–372 (1998).
20. Sampath, K. *et al.* Induction of the zebrafish ventral brain and floorplate requires cyclops/nodal signalling. *Nature* **395**, 185–189 (1998).
21. Feldman, B. *et al.* Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* **395**, 181–185 (1998).
22. Joseph, E. M. & Melton, D. A. *Xnr4* is a *Xenopus* nodal-related gene expressed in the Spemann organizer. *Dev. Biol.* **184**, 367–372 (1997).
23. Kimelman, D. & Griffin, K. J. Mesoderm induction: a postmodern view. *Cell* **94**, 419–421 (1998).
24. Zhang, J. *et al.* The role of maternal VegT in establishing the primary germ layers in *Xenopus* embryos. *Cell* **94**, 515–524 (1998).
25. Kofron, M. *et al.* Mesoderm induction in *Xenopus* is a zygotic event regulated by maternal VegT via TGF- β growth factors. *Development* **126**, 5759–5770 (1999).
26. Yasuo, H. & Lemaire, P. A two-step model for the fate determination of presumptive endodermal blastomeres in *Xenopus* embryos. *Curr. Biol.* **9**, 869–879 (1999).
27. Clements, D., Friday, R. V. & Woodland, H. R. Mode of action of VegT in mesoderm and endoderm formation. *Development* **126**, 4903–4911 (1999).
28. Shen, M. M., Wang, H. & Leder, P. A differential display strategy identifies *Cryptic*, a novel EGF-related gene expressed in the axial and lateral mesoderm during mouse gastrulation. *Development* **124**, 429–442 (1997).
29. Zhang, J., Talbot, W. S. & Schier, A. F. Positional cloning identifies zebrafish *one-eyed pinhead* as a permissive EGF-related ligand required during gastrulation. *Cell* **92**, 241–251 (1998).
30. Salomon, D. S., Bianco, C. & De Santis, M. *Cripto*: a novel epidermal growth factor (EGF)-related peptide in mammary gland development and neoplasia. *Bioessays* **21**, 61–70 (1999).
31. Gritsman, K. *et al.* The EGF-CFC protein *one-eyed pinhead* is essential for nodal signaling. *Cell* **97**, 121–132 (1999).
32. Ding, J. *et al.* *Cripto* is required for correct orientation of the anterior–posterior axis in the mouse embryo. *Nature* **395**, 702–707 (1998).
33. Schier, A. F., Neuhauss, S. C. F., Helde, K. A., Talbot, W. S. & Driever, W. The *one-eyed pinhead* gene functions in mesoderm and endoderm formation in zebrafish and interacts with *no tail*. *Development* **124**, 327–342 (1997).
34. Strahle, U. *et al.* *one-eyed pinhead* is required for development of the ventral midline of the zebrafish (*Danio rerio*) neural tube. *Genes Funct.* **1**, 131–148 (1997).
35. Meno, C. *et al.* Left–right asymmetric expression of the TGF- β -family member *lefty* in mouse embryos. *Nature* **381**, 151–155 (1996).
36. Meno, C. *et al.* Two closely-related left–right asymmetrically expressed genes, *lefty-1* and *lefty-2*: their distinct expression domains, chromosomal linkage and direct neuralizing activity in *Xenopus* embryos. *Genes Cell* **2**, 513–524 (1997).
37. Thisse, C. & Thisse, B. *Antivin*, a novel and divergent member of the TGF- β superfamily, negatively regulates mesoderm induction. *Development* **126**, 229–240 (1999).
38. Bisgrove, B. W., Essner, J. J. & Yost, H. J. Regulation of midline development by antagonism of *lefty* and *nodal* signaling. *Development* **126**, 3253–3262 (1999).
39. Cheng, A., Thisse, B., Thisse, C. & Wright, C. V. E. The *lefty*-related factor *Xatv* acts as a feedback inhibitor of Nodal signaling in mesoderm induction and L–R axis development in *Xenopus*. *Development* (in the press).
40. Bouwmeester, T., Kim, S., Sasai, Y., Lu, B. & De Robertis, E. M. *Cerberus* is a head-inducing secreted factor expressed in the anterior endoderm of Spemann's organizer. *Nature* **382**, 595–601 (1996).
41. Hsu, D. R., Economides, A. N., Wang, X., Eimon, P. M. & Harland, R. M. The *Xenopus* dorsalizing factor *Grenlin* identifies a novel family of secreted proteins that antagonize BMP activities. *Mol. Cell* **1**, 673–683 (1998).
42. Piccolo, S. *et al.* The head inducer *Cerberus* is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature* **397**, 707–710 (1999).
43. Rodriguez Esteban, C. *et al.* The novel *Cer*-like protein *Caronte* mediates the establishment of embryonic left–right asymmetry. *Nature* **401**, 243–251 (1999).
44. Simpson, E. H. *et al.* The mouse *Cer1* (*Cerberus* related or homologue) gene is not required for anterior pattern formation. *Dev. Biol.* **213**, 202–206 (1999).
45. Massagué, J. TGF- β signal transduction. *Annu. Rev. Biochem.* **67**, 753–791 (1998).
46. Whitman, M. Smads and early developmental signaling by the TGF- β superfamily. *Genes Dev.* **12**, 2445–2462 (1998).
47. Waldrup, W. R., Bikoff, E. K., Hoodless, P. A., Wrana, J. L. & Robertson, E. J. Smad2 signaling in extraembryonic tissues determines anterior–posterior polarity of the early mouse embryo. *Cell* **92**, 797–808 (1998).
48. Weinstein, M. *et al.* Failure of egg cylinder elongation and mesoderm induction in mouse embryos lacking the tumor suppressor *smad2*. *Proc. Natl Acad. Sci. USA* **95**, 9378–9383 (1998).
49. Nomura, M. & Li, E. Smad2 role in mesoderm formation, left–right patterning and craniofacial development. *Nature* **393**, 786–790 (1998).
50. Gu, Z. *et al.* The type I activin receptor *ActRIB* is required for egg cylinder organization and gastrulation in the mouse. *Genes Dev.* **12**, 844–857 (1998).
51. Song, J. *et al.* The type II activin receptors are essential for egg cylinder growth, gastrulation, and rostral head development in mice. *Dev. Biol.* **213**, 157–169 (1999).
52. Ramsdell, A. F. & Yost, H. J. Molecular mechanisms of vertebrate left–right development. *Trends Genet.* **14**, 459–465 (1998).
53. Yan, Y.-T. *et al.* Conserved requirement for *EGF-CFC* genes in vertebrate left–right axis formation. *Genes Dev.* **13**, 2527–2537 (1999).
54. Meno, C. *et al.* *lefty-1* is required for left–right determination as a regulator of *lefty-2* and nodal. *Cell* **94**, 287–297 (1998).
55. Oh, S. P. & Li, E. The signaling pathway mediated by the type IIB activin receptor controls axial patterning and lateral asymmetry in the mouse. *Genes Dev.* **11**, 1812–1826 (1997).
56. Gurdon, J. B., Dyson, S. & St Johnston, D. Cells' perception of position in a concentration gradient. *Cell* **95**, 159–162 (1998).
57. Gritsman, K., Talbot, W. S. & Schier, A. F. Nodal signaling patterns the organizer. *Development* (in the press).
58. Jones, C. M., Armes, N. & Smith, J. C. Signalling by TGF- β family members: short-range effects of *Xnr-2* and *BMP-4* contrast with the long-range effects of *activin*. *Curr. Biol.* **6**, 1468–1475 (1996).
59. Hatta, K., Kimmel, C. B., Ho, R. K. & Walker, C. The *cyclops* mutation blocks specification of the floor plate of the zebrafish central nervous system. *Nature* **350**, 339–341 (1991).
60. Heisenberg, C. P. & Nüsslein-Volhard, C. The function of *silberblick* in the positioning of the eye anlage in the zebrafish embryo. *Dev. Biol.* **184**, 85–94 (1997).
61. Matzuk, M. M., Kumar, T. R. & Bradley, A. Different phenotypes for mice deficient in either *activins* or *activin* receptor type II. *Nature* **374**, 356–360 (1995).
62. Sirard, C. *et al.* The tumor suppressor gene *Smad4/Dpc4* is required for gastrulation and later for anterior development of the mouse embryo. *Genes Dev.* **12**, 107–119 (1998).
63. Yang, X., Li, C., Xu, X. & Deng, C. The tumor suppressor *SMAD4/DPC4* is essential for epiblast proliferation and mesoderm induction in mice. *Proc. Natl Acad. Sci. USA* **95**, 3667–3672 (1998).
64. Yokouchi, Y., Vogan, K. J., Pearce, R. V. 2nd & Tabin, C. J. Antagonistic signaling by *Caronte*, a novel *Cerberus*-related gene, establishes left–right asymmetric gene expression. *Cell* **98**, 573–583 (1999).
65. Zhu, L. *et al.* *Cerberus* regulates left–right asymmetry of the embryonic head and heart. *Curr. Biol.* **9**, 931–938 (1999).
66. Rodaway, A. *et al.* Induction of the mesendoderm in the zebrafish germ ring by yolk cell-derived TGF- β family signals and discrimination of mesoderm and endoderm by FGF. *Development* **126**, 3067–3078 (1999).
67. Thisse, B., Wright, C. V. E. & Thisse, B. *Activin- and Nodal-related factors control antero–posterior patterning of the zebrafish embryo*. *Nature* **403**, 425–428 (2000).
68. Varlet, L., Collignon, J. & Robertson, E. J. *nodal* expression in the primitive endoderm is required for specification of the anterior axis during mouse gastrulation. *Development* **124**, 1033–1044 (1997).
69. Osada, S. I. & Wright, C. V. *Xenopus* nodal-related signaling is essential for mesendodermal patterning during early embryogenesis. *Development* **126**, 3229–3240 (1999).
70. Beddington, R. S. P. & Robertson, E. J. Anterior patterning in mouse. *Trends Genet.* **14**, 277–284 (1998).
71. Schuurte, H. E. *et al.* Sonic hedgehog is not required for the induction of medial floor plate cells in the zebrafish. *Development* **125**, 2983–2993 (1998).

Acknowledgements

We thank C. Abate-Shen, R. Burdine, K. Gritsman, D. Kimelman, D. Reinberg, D. Rifkin, D. Ron, W. Talbot, E. White and A. Zychlinski for comments on the manuscript. Only a partial reference list is included owing to space constraints. The authors are supported by grants from the NIH (A.F.S., M.M.S.) and the US Army Breast Cancer Research Program (M.M.S.). A.F.S. is a Scholar of the McKnight Endowment Fund for Neuroscience.

Correspondence should be addressed to A.F.S. (e-mail: schier@saturn.med.nyu.edu) or M.M.S. (e-mail: mshen@cabm.rutgers.edu).

A pigment-binding protein essential for regulation of photosynthetic light harvesting

Xiao-Ping Li*, Olle Björkman†, Connie Shih†, Arthur R. Grossman†, Magnus Rosenquist‡§, Stefan Jansson‡ & Krishna K. Niyogi*

* Department of Plant and Microbial Biology, University of California, Berkeley, California 94720-3102, USA

† Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA

‡ Umeå Plant Science Center, Department of Plant Physiology, University of Umeå, 901 87 Umeå, Sweden

Photosynthetic light harvesting in plants is regulated in response to changes in incident light intensity. Absorption of light that exceeds a plant's capacity for fixation of CO₂ results in thermal dissipation of excitation energy in the pigment antenna of photosystem II by a poorly understood mechanism. This regulatory process, termed nonphotochemical quenching, maintains the balance between dissipation and utilization of light energy to minimize generation of oxidizing molecules, thereby protecting the plant against photo-oxidative damage. To identify specific proteins that are involved in nonphotochemical quenching, we have isolated mutants of *Arabidopsis thaliana* that cannot dissipate excess absorbed light energy. Here we show that the gene encoding PsbS, an intrinsic chlorophyll-binding protein of photosystem II, is necessary for nonphotochemical quenching but not for efficient light harvesting and photosynthesis. These results indicate that PsbS may be the site for nonphotochemical quenching, a finding that has implications for the functional evolution of pigment-binding proteins.

Photosynthetic reactions in plants convert light energy into chemical energy that supports much of the life on Earth. However, the involvement of highly reactive intermediates necessitates regulation of photosynthesis to cope with the constantly changing quantity of light in natural environments and to protect against photo-oxidative damage^{1–4}. Photosynthetic light harvesting in plants is subject to feedback regulation by the changes in pH (Δ pH) generated by electron transport in the chloroplast³. Whenever light absorption exceeds a plant's capacity for CO₂ fixation, accumulation of protons in the thylakoid lumen results in thermal dissipation of absorbed light energy in the pigment antenna of photosystem II. Absorption of excess sunlight occurs daily for many photosynthetic organisms, especially when growth is limited by environmental stresses such as extremes of temperature, drought, salinity and nutrient deprivation.

Photoprotective thermal dissipation of absorbed light energy is measured as nonphotochemical quenching of chlorophyll fluorescence (NPQ) during exposure of plants to high light³. Although several processes can contribute to NPQ, the major component in wild-type *Arabidopsis* and other plants is Δ pH-dependent NPQ, referred to as qE, which is characterized by rapid induction and relaxation kinetics (Fig. 1a, b). During induction of qE by light, the build-up of Δ pH activates synthesis of a specific xanthophyll pigment, zeaxanthin, through a xanthophyll cycle⁵. Allosteric binding of both H⁺ and zeaxanthin to light-harvesting antenna proteins is suggested to cause a conformational change that leads to quenching of excitation energy^{3,6,7}. The rapid relaxation of qE during subsequent darkness (Fig. 1a, b) reflects the decay of the light-induced Δ pH. Analysis of a mutant that is defective in the xanthophyll cycle, *npq1*, showed that zeaxanthin synthesis is necessary for about 70% of the total NPQ and 80% of the qE in *Arabidopsis*⁸ (Fig. 1b). However, zeaxanthin is not sufficient for qE in the absence of the Δ pH, as shown using *npq2* (*aba1*) mutants that accumulate zeaxanthin constitutively^{8–10}.

Although qE has been the subject of intense research for more than a decade, its exact site and molecular mechanism are unknown.

The possible involvement of specific proteins in the light-harvesting antenna of photosystem II has been addressed by identification of proteins that bind zeaxanthin¹¹ and dicyclohexylcarbodiimide (DCCD, an inhibitor of qE *in vitro*)¹² and by characterization of plants that are deficient in various antenna proteins^{13,14}. These experiments have led to the suggestion that qE occurs in the Lhcb4 (CP29) and Lhcb5 (CP26) antenna proteins^{3,7,11,15,16}, but there has been no direct evidence for the function of any specific protein in qE.

npq4 is defective in nonphotochemical quenching

To identify components in addition to the Δ pH and xanthophylls that are necessary for qE, we used a chlorophyll fluorescence video-imaging system⁸ to isolate *Arabidopsis* mutants with altered NPQ but normal pigment composition. Measurement of NPQ induction and relaxation kinetics showed that the *npq4-1* mutant was defective specifically in the qE component of NPQ (Fig. 1a, b). However, the *npq4-1* mutant exhibited zeaxanthin synthesis in high light that was indistinguishable in extent (Table 1) and kinetics (data not shown) from that of the wild type, in contrast to the previously characterized *npq1* mutants⁸.

Genetic analysis showed that the phenotype of *npq4-1* plants was due to a single, semidominant, nuclear mutation (Table 2). The Npq[–] phenotype of an *npq4-1 npq1-2* double mutant was indistinguishable from that of *npq4-1* (Fig. 1b), showing that the *npq4-1*

Table 1 Xanthophyll cycle pigments in wild-type and *npq4-1* leaves

		mmol per mol chlorophyll a				
		V	A	Z	V + A + Z	(A + Z)/(V + A + Z)
Wild type	before HL	53 ± 8	16 ± 2	13 ± 1	82 ± 9	0.36 ± 0.03
	after HL	18 ± 1	8 ± 1	43 ± 9	70 ± 10	0.73 ± 0.03
<i>npq4-1</i>	before HL	54 ± 2	12 ± 1	6 ± 1	73 ± 2	0.26 ± 0.02
	after HL	17 ± 1	8 ± 1	43 ± 5	68 ± 6	0.75 ± 0.02

Pigment composition of leaves was measured by high-performance liquid chromatography either before or after exposure to high light (HL, 640 μ mol photons m^{–2} sec^{–1}) for 30 min. Values are means $\pm$ s.e. (*n* = 4).

§ Present address: Plant Biochemistry, University of Lund, PO Box 117, S-221 00 Lund, Sweden.

mutation blocks even the small amount of qE that is detected in *npq1-2* plants.

npq4 lacks the conformational change

The *npq4-1* mutant lacks the Δ pH- and zeaxanthin-dependent conformational change in the thylakoid membrane that is necessary for qE. This conformational change is monitored by a light-induced change in absorbance at 535 nm (ΔA_{535})^{17,18}, which appears in wild-type leaves as a shoulder on a major absorbance change at 505 nm (Fig. 2a). The ΔA_{505} , which is attributable to conversion of violaxanthin to zeaxanthin through the xanthophyll cycle¹⁹, occurred in the leaves of both wild-type and *npq4-1* plants, consistent with the normal synthesis of zeaxanthin that was observed in *npq4-1* (Table 1). However, the ΔA_{535} was absent in the *npq4-1* mutant, as demonstrated by the lack of the shoulder on the ΔA_{505} peak (Fig. 2a).

To determine whether *npq4-1* retained any ΔA_{535} that had been obscured by the major ΔA_{505} signal, we examined an *npq4-1 npq2-1* double mutant. Plants carrying the *npq2-1* mutation accumulate zeaxanthin constitutively and lack the ΔA_{505} owing to the absence of xanthophyll cycling⁸, allowing unambiguous visualization of the

ΔA_{535} (Fig. 2b). In contrast to the *npq2-1* single mutant, the *npq4-1 npq2-1* double mutant completely lacked a detectable ΔA_{535} signal (Fig. 2b). Together, these results indicate that the *npq4-1* mutation may define a gene that is necessary for both the conformational change in the antenna and qE.

The *NPQ4* gene encodes PsbS

To isolate the *NPQ4* gene, we used a combination of molecular and visible genetic markers to map the *npq4-1* mutation to chromosome 1, ~0.7 cM south of the *chl1* locus (Fig. 3a). Hybridization with a *CH1* complementary DNA²⁰ identified several bacterial artificial chromosome (BAC) clones, some of which also contained the *Arabidopsis* gene encoding PsbS²¹, an intrinsic pigment-binding photosystem II subunit (also known as CP22) of previously unknown function²². Polymerase chain reaction (PCR) amplification of the *psbS* gene from the *npq4-1* mutant was unsuccessful (data not shown), indicating that the fast neutron-induced *npq4-1* allele may contain a DNA rearrangement or deletion affecting *psbS*. DNA gel blot analysis confirmed that the *psbS* gene is completely absent from the genome of *npq4-1* (Fig. 3b). Four other alleles of *npq4*, which were isolated following mutagenesis with ethylmethane sulphonate, contained missense mutations in *psbS* (Fig. 3c). Complementation of the *npq4-1* (Fig. 1c) and *npq4-4* mutations (data not shown) by transformation with a wild-type copy of the *psbS* gene verified that the *NPQ4* gene is *psbS*.

Light-harvesting function in *npq4*

The PsbS protein is a member of the chlorophyll *a/b*-binding, light-harvesting complex (LHC) family of proteins^{21,23,24}. However, PsbS is present in many oxygen-evolving photosystem II preparations²⁵ that are depleted of typical LHC proteins, indicating that PsbS may be closely associated with the photosystem II reaction centre,

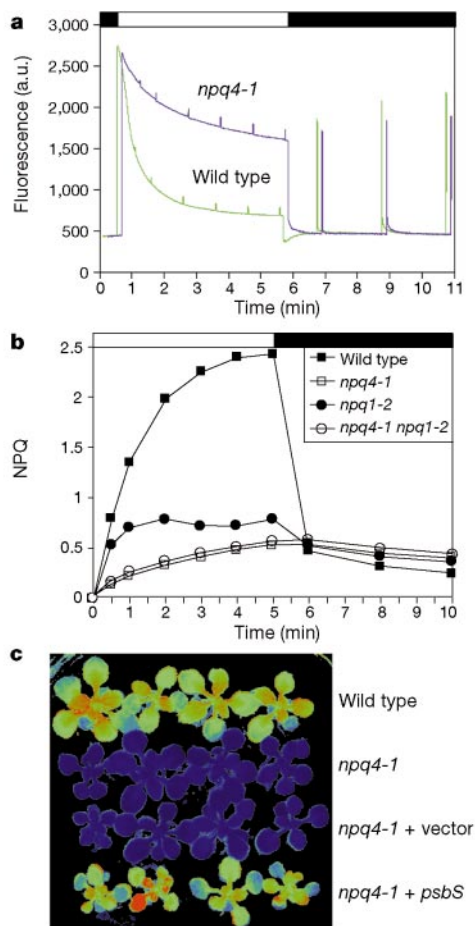


Figure 1 Nonphotochemical quenching phenotypes. White bars above graphs indicate periods of illumination with high light ($1250 \mu\text{mol photons m}^{-2} \text{sec}^{-1}$); black bars indicate darkness (with only very weak measuring light). **a**, Chlorophyll fluorescence during induction and relaxation of NPQ in wild type and *npq4-1*. The *npq4-1* trace is offset by 10 s relative to that of the wild type. **b**, NPQ calculated from fluorescence data for wild type, *npq4-1*, *npq1-2* and an *npq4-1 npq1-2* double mutant. **c**, Image of NPQ occurring after 1 min of illumination with $800 \mu\text{mol photons m}^{-2} \text{sec}^{-1}$. *npq4-1* + vector, *npq4-1* plants transformed with pPZP121; *npq4-1* + *psbS*, *npq4-1* plants transformed with pXPL1. a.u., arbitrary units.

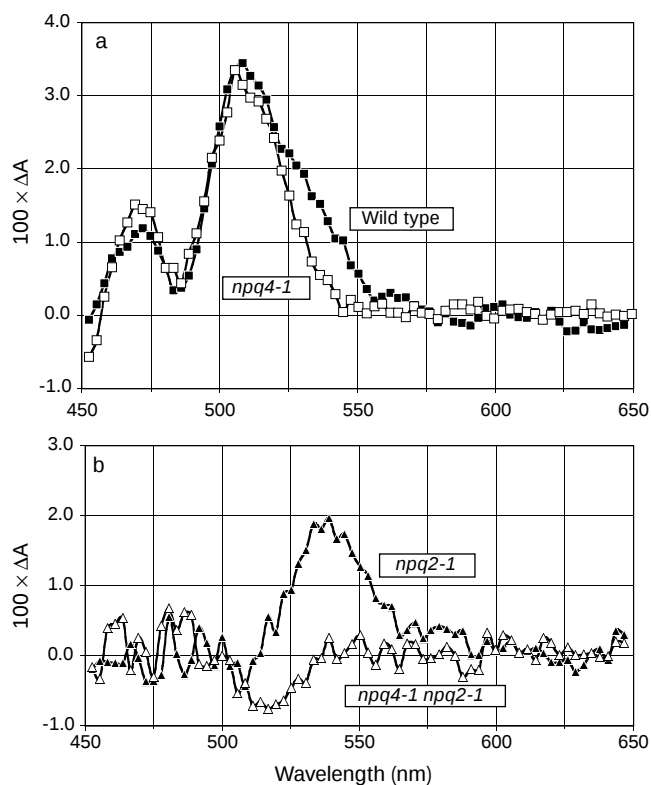


Figure 2 Light-induced spectral absorbance changes in leaves. **a**, Wild type (filled squares) and *npq4-1* (open squares). **b**, *npq2-1* (filled triangles) and *npq4-1 npq2-1* (open triangles).

perhaps at the interface between the reaction centre core and the peripheral light-harvesting antenna²⁶. Despite the absence of the PsbS protein in the *npq4-1* deletion mutant (Fig. 4), light harvesting and photosynthesis were not obviously affected. Both the quantum yield and the maximum rate of photosynthetic oxygen evolution in the *npq4-1* mutant were the same as those of the wild type (Table 3). The maximum quantum yield of photosystem II electron transport, determined from chlorophyll fluorescence measurements at room temperature, was similarly unaffected in *npq4-1* plants (Table 3). In addition, the fluorescence emission spectra of *npq4-1* and wild-type thylakoids at 77K were identical (data not shown). Growth of *npq4-1* plants at limiting light intensities was indistinguishable from that of the wild type (data not shown). The other LHC proteins that comprise the peripheral light-harvesting antenna of photosystem II (Lhcb1-Lhcb6) were all present at wild-type levels in *npq4-1* (Fig. 4).

Discussion

These results show clearly that the PsbS protein contributes to photoprotective energy dissipation rather than photosynthetic light harvesting. Although the exact role of PsbS in qE remains to be determined, the lack of both qE (Fig. 1) and the protonation-induced conformational change (ΔA_{535}) (Fig. 2) in the *npq4-1*

mutant indicates that binding of one or more protons by PsbS may be a necessary feature of the qE mechanism. Several acidic amino-acid residues in PsbS on the lumen side of the thylakoid membrane are candidate proton-binding sites (Fig. 3c).

With the observation that the isolated PsbS protein binds chlorophylls and xanthophylls²², our results suggest a model in which PsbS, rather than other pigment-binding LHC proteins in the antenna of photosystem II^{3,7,11,15,16}, is the site of ΔpH - and xanthophyll-dependent excitation quenching. Consistent with this model, plants that are deficient in LHC proteins but retain qE have wild-type levels of PsbS^{27,28}. Upon protonation of PsbS, a conformational change could allow for quenching of single excited chlorophyll either by direct energy transfer from chlorophyll to zeaxanthin^{5,29} bound to PsbS or by a zeaxanthin-dependent chlorophyll–chlorophyll^{3,16} or chlorophyll–protein interaction occurring within PsbS. Alternatively, protonation of PsbS may be necessary for conformational changes and quenching that actually occur in adjacent LHC proteins of an antenna subcomplex. Indeed, isolated Lhcb complexes have been observed to exhibit pH-dependent quenching of chlorophyll fluorescence *in vitro* that is modulated by factors, such as zeaxanthin, that affect qE *in vivo*^{30–32}. It is also possible that PsbS is simply necessary to maintain LHC proteins in a proper

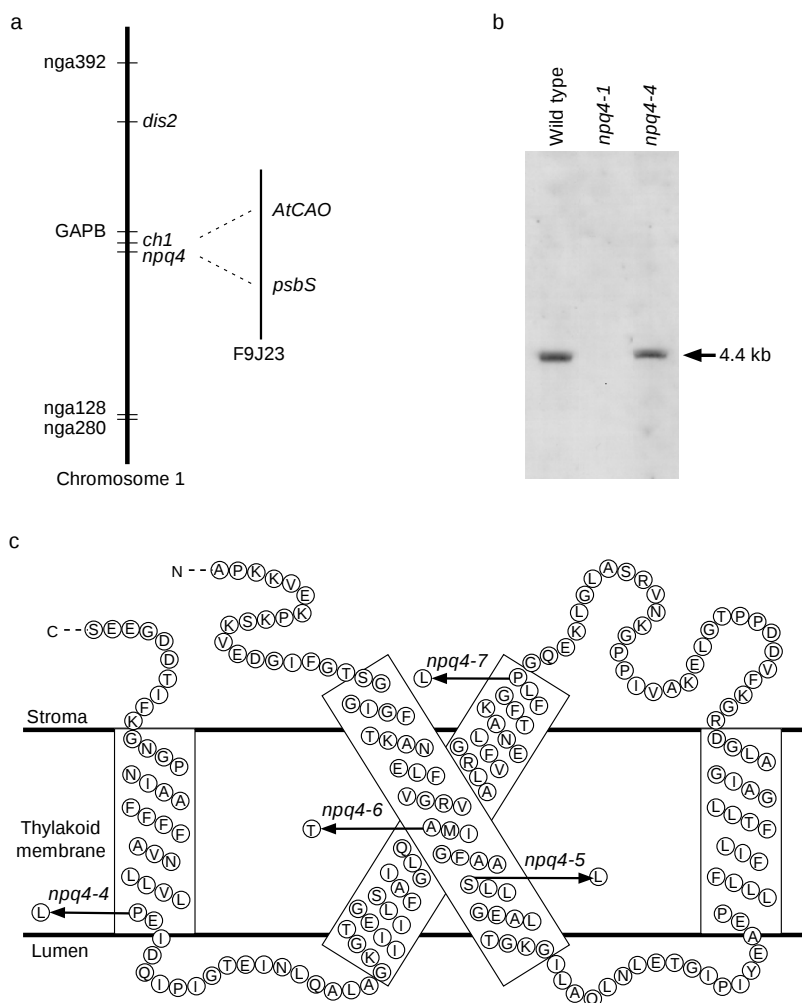


Figure 3 Cloning of *NPQ4*. **a**, Map position of *npq4-1*. Molecular markers are shown to the left of the vertical line that represents part of chromosome 1; visible markers are shown on the right. F9J23 is a BAC clone containing both *CH1* (*AtCaO*, ref. 20) and *psbS*. The relative positions of *AtCaO* and *psbS* are shown, but their exact locations within F9J23 were not determined. **b**, DNA gel blot analysis. Genomic DNA from wild type, *npq4-1* and

npq4-4 was digested with *Xba*I and hybridized with a *psbS* probe. **c**, Schematic representation of the PsbS protein. The positions of missense mutations in *npq4* point mutant alleles are indicated. The predicted topology of the PsbS protein is based on the crystal structure of the related LHC-II protein⁴⁷.

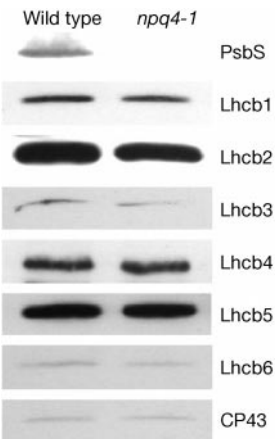


Figure 4 LHC protein levels in wild type and *npq4-1*. Thylakoid protein samples corresponding to equal amounts of chlorophyll were separated by SDS-PAGE, and immunoblot analysis was performed with polyclonal antibodies directed against each of the indicated photosystem II proteins. Lhcb1, Lhcb2 and Lhcb3 are components of trimers

in the most peripheral antenna of photosystem II; Lhcb4, Lhcb5 and Lhcb6 are monomeric, peripheral antenna proteins (also known as CP29, CP26 and CP24, respectively)²¹. CP43 is a core antenna subunit of photosystem II.

supramolecular organization that allows qE to occur. However, efficient photosynthetic light harvesting by *npq4-1* in limiting light (Table 3) and accumulation of wild-type levels of other LHC proteins in *npq4-1* (Fig. 4) indicate that the qE defect of the mutant is not due to an obvious indirect effect on the photosystem II antenna. Further structural analysis of the PsbS protein may help to define possible proton- and/or zeaxanthin-binding sites and to elucidate the molecular mechanism of ΔpH- and xanthophyll-dependent regulation of light harvesting.

The semidominance of the loss-of-function *npq4-1* mutation (Table 2) indicates that the *psbS* gene dosage may be a critical determinant of qE capacity. The maximum extent of qE is considered to be an important factor for adaptation of plants to adverse environments, and plants growing in excessive light generally exhibit greater maximal qE than shade plants^{33–35}. In *Arabidopsis*, the level of *psbS* messenger RNA is increased severalfold in response to high light (unpublished results), consistent with a possible relationship between *psbS* expression and qE capacity. Accumulation of PsbS protein in etiolated seedlings of spinach²⁷ may allow photoprotective qE upon exposure to light during the early stages of greening. In overwintering pine trees, which exhibit high levels of slowly reversible NPQ, increases in the amount of PsbS are associated with a reorganization of pigment-binding proteins³⁶, suggesting that PsbS is involved in more than one type of NPQ.

Unlike typical LHC proteins with three transmembrane α-helices, the PsbS protein has four helices that span the thylakoid membrane²⁶ (Fig. 3c). The first and third helices of PsbS are similar to each other and to the first and third helices of three-helix LHC proteins. In addition, the second and fourth helices of PsbS are similar to each other and to the second helix of three-helix LHC proteins, indicating that PsbS may have arisen by internal duplica-

tion of a gene encoding a two-helix protein and that the three-helix LHC proteins then evolved from a four-helix ancestor like PsbS^{23,24,37}. Genes encoding one-helix proteins (similar to the first and third helices of PsbS), as well as two-helix proteins that resemble proposed intermediate forms in the evolution of LHC proteins, have been identified^{21,38}. Expression of several genes encoding these ‘ancestral’ types of LHC protein is induced by high light stress^{21,38}, suggesting that they are photoprotective. Our finding that PsbS is involved in energy dissipation rather than light harvesting supports the hypothesis that, during the evolution of oxygenic photosynthesis, members of the LHC protein family with roles in photoprotection appeared before those involved in light harvesting^{39,40}. □

Methods

Mutant isolation

Mutants of *Arabidopsis thaliana* (ecotype Col-0) were identified by chlorophyll fluorescence video imaging⁸ of M₂ seedlings derived from mutagenesis with fast-neutron bombardment (Lehle Seeds) or 0.3% (v/v) ethylmethane sulphonate. Plants were grown on minimal agar medium in petri plates or in Sunshine growth mix (Sun Gro Horticulture) in pots in a greenhouse⁸.

Chlorophyll fluorescence, spectroscopy and oxygen evolution

We measured chlorophyll fluorescence from attached rosette leaves at room temperature with an FMS2 fluorometer (Hansatech). NPQ was calculated as $(F_m - F'_m)/F'_m$. Measurements of light-induced spectral absorbance changes induced by illumination with 1,900 μmol photons m⁻² sec⁻¹ for 3 min were performed as described⁸. Photosynthetic oxygen evolution was measured as a function of incident photon flux density with a leaf disc electrode (LD2/2, Hansatech).

Pigment analysis

Chlorophylls and carotenoids were analysed by high-performance liquid chromatography (HP1100, Hewlett-Packard) on a Microsorb-MV column as described⁴¹. Carotenoids were quantified using standard curves of purified pigments (VKI).

Table 2 Results of cross between *npq4-1* and wild-type plants

Cross	Type	Total	Npq ⁺	Npq ^{+/-}	Npq ⁻
<i>npq4-1</i> / <i>npq4-1</i> × NPQ4/NPQ4	F ₁	8	0	8	0
	F ₂	233	60	106	67

Pollen from the male parent (listed first) was crossed onto stigmas of the female parent to generate F₁ seeds. The F₁ plants were allowed to self-pollinate to generate F₂ seeds. The phenotypes of F₁ and F₂ plants were scored by fluorescence video imaging after 12 days of growth on agar medium. The heterozygous F₁ plants had an intermediate phenotype (Npq^{+/-}). The 1:2:1 segregation of the Npq phenotype in the F₂ generation is consistent with the hypothesis that *npq4-1* is a single semidominant, nuclear mutation ($\chi^2 = 2.3$, $P > 0.1$).

Table 3 Photosynthetic parameters of wild-type and *npq4-1* leaves

Parameter	Wild type	<i>npq4-1</i>
$\Phi(O_2)$	0.0818 ± 0.0024 (n = 5)	0.0812 ± 0.0033 (n = 5)
P_{max}	18.15 ± 1.06 (n = 5)	18.87 ± 0.92 (n = 5)
F_v/F_m	0.837 ± 0.003 (n = 8)	0.836 ± 0.004 (n = 8)

$\Phi(O_2)$, apparent quantum yield of O₂ evolution (O₂ evolved per incident photon); P_{max} , maximum rate of O₂ evolution (μmol O₂ m⁻² sec⁻¹); F_v/F_m , maximum quantum yield of photosystem II electron transport. Values are means ± s.e.

Genetic mapping and molecular biology

The *npq4-1* mutation was mapped initially by scoring PCR-based markers^{42,43} on *npq4-1/npq4-1* F₂ progeny derived from a cross between *npq4-1/npq4-1* (Col-0 ecotype) and *NPQ4/NPQ4* (Ler-0 ecotype). Recombination fractions (number of crossovers/number of chromosomes scored) were 6/40, 8/44, 9/42 and 1/48 between *npq4-1* and *nga392*, *nga128*, *nga280* and *GAPB*, respectively. Among 2,072 F₂ progeny of a cross between *npq4-1/npq4-1* and (*dis2-1 ch1-1*)/(*dis2-1 ch1-1*), we detected seven (*DIS2 CH1 NPQ4*)/(*dis2-1 ch1-1 NPQ4*) recombinant chromosomes. We used a *CH1* cDNA clone (103D24T7) as a hybridization probe (AlkPhos Direct, Amersham Pharmacia Biotech) to identify BAC clones containing the *CH1* gene. The *psbS* gene was amplified by PCR from BAC clones and genomic DNA of wild-type and *npq4* mutants using primers KN118 (5'-TCCTTCTCTCATCCTCAGAAA-3') and KN119 (5'-CAACATGAAGAGAAGGTCACA-3'), and 1.5-kilobase (kb) PCR products were used as hybridization probes or for DNA sequencing. For complementation of *npq4-1* and *npq4-4*, a 4.4-kb *XbaI* fragment of BAC clone F9J23 containing the *psbS* gene was subcloned into the pPZP121 vector⁴⁴ to generate pXPL1. Plants were transformed⁴⁵ using *Agrobacterium tumefaciens* GV3101 containing either pPZP121 or pXPL1.

Immunoblot analysis

Thylakoid membranes were prepared and analysed by SDS-PAGE and immunoblotting using monospecific antibodies⁴⁶.

Received 12 October; accepted 6 December 1999.

- Demmig-Adams, B. & Adams, W. W. III. Photoprotection and other responses of plants to high light stress. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43**, 599–626 (1992).
- Dau, H. Short-term adaptation of plants to changing light intensities and its relation to photosystem II photochemistry and fluorescence emission. *J. Photochem. Photobiol. B: Biol.* **26**, 3–27 (1994).
- Horton, P., Ruban, A. V. & Walters, R. G. Regulation of light harvesting in green plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 655–684 (1996).
- Niyogi, K. K. Photoprotection revisited: Genetic and molecular approaches. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 333–359 (1999).
- Demmig-Adams, B. Carotenoids and photoprotection in plants: A role for the xanthophyll zeaxanthin. *Biochim. Biophys. Acta* **1020**, 1–24 (1990).
- Horton, P. et al. Control of the light-harvesting function of chloroplast membranes by aggregation of the LHCII chlorophyll–protein complex. *FEBS Lett.* **292**, 1–4 (1991).
- Gilmore, A. M. Mechanistic aspects of xanthophyll cycle-dependent photoprotection in higher plant chloroplasts and leaves. *Physiol. Plant.* **99**, 197–209 (1997).
- Niyogi, K. K., Grossman, A. R. & Björkman, O. Arabidopsis mutants define a central role for the xanthophyll cycle in the regulation of photosynthetic energy conversion. *Plant Cell* **10**, 1121–1134 (1998).
- Tardy, F. & Havaux, M. Photosynthesis, chlorophyll fluorescence, light-harvesting system and photoinhibition resistance of a zeaxanthin-accumulating mutant of *Arabidopsis thaliana*. *J. Photochem. Photobiol. B: Biol.* **34**, 87–94 (1996).
- Hurry, V., Anderson, J. M., Chow, W. S. & Osmond, C. B. Accumulation of zeaxanthin in abscisic acid-deficient mutants of *Arabidopsis* does not affect chlorophyll fluorescence quenching or sensitivity to photoinhibition *in vivo*. *Plant Physiol.* **113**, 639–648 (1997).
- Bassi, R., Pineau, B., Dainese, P. & Marquardt, J. Carotenoid-binding proteins of photosystem II. *Eur. J. Biochem.* **212**, 297–303 (1993).
- Walters, R. G., Ruban, A. V. & Horton, P. Higher plant light-harvesting complexes LHCIIa and LHCIIc are bound by dicyclohexylcarbodiimide during inhibition of energy dissipation. *Eur. J. Biochem.* **226**, 1063–1069 (1994).
- Jahns, P. & Krause, G. H. Xanthophyll cycle and energy-dependent fluorescence quenching in leaves from pea plants grown under intermittent light. *Planta* **192**, 176–182 (1994).
- Gilmore, A. M., Hazlett, T. L., Debrunner, P. G. & Govindjee. Photosystem II chlorophyll *a* fluorescence lifetimes and intensity are independent of the antenna size differences between barley wild-type and *chlorina* mutants: Photochemical quenching and xanthophyll cycle-dependent non-photochemical quenching of fluorescence. *Photosynth. Res.* **48**, 171–187 (1996).
- Horton, P. & Ruban, A. V. Regulation of Photosystem II. *Photosynth. Res.* **34**, 375–385 (1992).
- Crofts, A. R. & Yerkes, C. T. A molecular mechanism for q_E-quenching. *FEBS Lett.* **352**, 265–270 (1994).
- Ruban, A. V., Young, A. J. & Horton, P. Induction of nonphotochemical energy dissipation and absorbance changes in leaves. Evidence for changes in state of the light-harvesting system of photosystem II *in vivo*. *Plant Physiol.* **102**, 741–750 (1993).
- Bilger, W. & Björkman, O. Relationships among violaxanthin deepoxidation, thylakoid membrane conformation, and nonphotochemical chlorophyll fluorescence quenching in leaves of cotton (*Gossypium hirsutum* L.). *Planta* **193**, 238–246 (1994).
- Yamamoto, H. Y. & Kamite, L. The effects of dithiothreitol on violaxanthin de-epoxidation and absorbance changes in the 500-nm region. *Biochim. Biophys. Acta* **267**, 538–543 (1972).
- Espineda, C. E., Linford, A. S., Devine, D. & Brusslan, J. A. The *AtCAO* gene, encoding chlorophyll *a* oxygenase, is required for chlorophyll *b* synthesis in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **96**, 10507–10511 (1999).
- Jansson, S. A guide to the Lhc genes and their relatives in *Arabidopsis*. *Trends Plant Sci.* **4**, 236–240 (1999).

- Funk, C. et al. The PSII-S protein of higher plants: A new type of pigment-binding protein. *Biochemistry* **34**, 11133–11141 (1995).
- Kim, S. et al. Characterization of a spinach *psbS* cDNA encoding the 22 kDa protein of photosystem II. *FEBS Lett.* **314**, 67–71 (1992).
- Wedel, N., Klein, R., Ljungberg, U., Andersson, B. & Herrmann, R. G. The single-copy gene *psbS* codes for a phylogenetically intriguing 22 kDa polypeptide of photosystem II. *FEBS Lett.* **314**, 61–66 (1992).
- Ghanotakis, D. F. et al. Comparative structural and catalytic properties of oxygen-evolving photosystem II preparations. *Photosynth. Res.* **14**, 191–199 (1987).
- Kim, S., Pichersky, E. & Yocum, C. F. Topological studies of spinach 22 kDa protein of Photosystem II. *Biochim. Biophys. Acta* **1188**, 339–348 (1994).
- Funk, C., Adamska, I., Green, B. R., Andersson, B. & Renger, G. The nuclear-encoded chlorophyll-binding photosystem II-S protein is stable in the absence of pigments. *J. Biol. Chem.* **270**, 30141–30147 (1995).
- Bossmann, B., Knoetzel, J. & Jansson, S. Screening of *chlorina* mutants of barley (*Hordeum vulgare* L.) with antibodies against light-harvesting proteins of PS I and PS II: Absence of specific antenna proteins. *Photosynth. Res.* **52**, 127–136 (1999).
- Owens, T. G. in *Photoinhibition of Photosynthesis: From Molecular Mechanisms to the Field* (eds Baker, N. R. & Bowyer, J. R.) 95–109 (BIOS Scientific, Oxford, 1994).
- Phillip, D., Ruban, A. V., Horton, P., Asato, A. & Young, A. J. Quenching of chlorophyll fluorescence in the major light-harvesting complex of photosystem II: a systematic study of the effect of carotenoid structure. *Proc. Natl Acad. Sci. USA* **93**, 1492–1497 (1996).
- Ruban, A. V., Young, A. J. & Horton, P. Dynamic properties of the minor chlorophyll *a/b* binding proteins of photosystem II, an *in vitro* model for photoprotective energy dissipation in the photosynthetic membrane of green plants. *Biochemistry* **35**, 674–678 (1996).
- Ruban, A. V. & Horton, P. The xanthophyll cycle modulates the kinetics of nonphotochemical energy dissipation in isolated light-harvesting complexes, intact chloroplasts, and leaves of spinach. *Plant Physiol.* **119**, 531–542 (1999).
- Johnson, G. N., Young, A. J., Scholes, J. D. & Horton, P. The dissipation of excess excitation energy in British plant species. *Plant Cell Environ.* **16**, 673–679 (1993).
- Demmig-Adams, B. & Adams, W. W., III. Capacity for energy dissipation in the pigment bed in leaves with different xanthophyll cycle pools. *Aust. J. Plant Physiol.* **21**, 575–588 (1994).
- Demmig-Adams, B. Survey of thermal energy dissipation and pigment composition in sun and shade leaves. *Plant Cell Physiol.* **39**, 474–482 (1998).
- Ottander, C., Campbell, D. & Öquist, G. Seasonal changes in photosystem II organisation and pigment composition in *Pinus sylvestris*. *Planta* **197**, 176–183 (1995).
- Green, B. R. & Pichersky, E. Hypothesis for the evolution of three-helix Chl *a/b* and Chl *a/c* light-harvesting proteins from two-helix and four-helix ancestors. *Photosynth. Res.* **39**, 149–162 (1994).
- Dolganov, N. A. M., Bhaya, D. & Grossman, A. R. Cyanobacterial protein with similarity to the chlorophyll *a/b* binding proteins of higher plants: Evolution and regulation. *Proc. Natl Acad. Sci. USA* **92**, 636–640 (1995).
- Green, B. R. & Kühlbrandt, W. Sequence conservation of light-harvesting and stress-response proteins in relation to the three-dimensional molecular structure of LHCII. *Photosynth. Res.* **44**, 139–148 (1995).
- Douglas, S. E. Plastid evolution: Origins, diversity, trends. *Curr. Op. Genet. Dev.* **8**, 655–661 (1998).
- Niyogi, K. K., Björkman, O. & Grossman, A. R. The roles of specific xanthophylls in photoprotection. *Proc. Natl Acad. Sci. USA* **94**, 14162–14167 (1997).
- Bell, C. J. & Ecker, J. R. Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. *Genomics* **19**, 137–144 (1994).
- Konieczny, A. & Ausubel, F. M. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* **4**, 403–410 (1993).
- Hajdukiewicz, P., Svab, Z. & Maliga, P. The small, versatile pPZP family of *Agrobacterium* binary vectors for plant transformation. *Plant Mol. Biol.* **25**, 989–994 (1994).
- Clough, S. J. & Bent, A. F. Floral dip: A simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
- Jansson, S., Stefánsson, H., Nyström, U., Gustafsson, P. & Albertsson, P.-A. Antenna protein composition of PS I and PS II in thylakoid sub-domains. *Biochim. Biophys. Acta* **1320**, 297–309 (1997).
- Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367**, 614–621 (1994).

Acknowledgements

We thank A. K. Tran and V. Canale for technical assistance; J. Brusslan for unpublished data on *CHI*; T. Shikanai for the *npq4-4* allele; C. Funk, J. Knötzel and A. Staehelin for antibodies; R. Malkin for comments on the manuscript; and the Arabidopsis Biological Resource Center for strains and DNA clones. This work was supported by grants from the U.S. Department of Agriculture–National Research Initiative Competitive Grants Program and the Searle Scholars Program/The Chicago Community Trust to K.K.N., a grant from the National Science Foundation to A.G. and O.B., and grants from the Swedish Forestry and Agricultural Research Council and the Foundation for Strategic Research to S.J. When this work was initiated, K.K.N. was supported as a Department of Energy Biosciences Fellow of the Life Sciences Research Foundation.

Correspondence and requests for materials should be addressed to K.K.N. (e-mail: niyogi@nature.berkeley.edu).

Probing bulk states of correlated electron systems by high-resolution resonance photoemission

A. Sekiyama*, T. Iwasaki*, K. Matsuda*, Y. Saitoh†, Y. Ônuki‡ & S. Suga*

* Department of Material Physics, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan

† Department of Synchrotron Research, Kansai Research Establishment, Japan Atomic Energy Research Institute, SPring-8, Mikazuki, Hyogo 679-5148, Japan

‡ Department of Physics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan

Electron correlations are known to play an important role in determining the unusual physical properties of a variety of compounds. Such properties include high-temperature superconductivity¹, heavy fermion behaviour^{2,3} and metal-to-insulator transitions^{4,5}. High-resolution photoelectron spectroscopy (PES) provides a means of directly probing the electronic states (particularly those near the Fermi level) in these materials, but the short photoelectron mean free paths⁶ (≤ 5 Å) associated with the low excitation energies conventionally used (≤ 120 eV) make this a surface-sensitive technique. Now that high-resolution PES is possible at much higher energies⁷, with mean free paths as long as 15 Å (ref. 6), it should become feasible to probe the bulk electronic states in these materials. Here we demonstrate the power of this technique by applying it to the cerium compounds CeRu₂Si₂ and CeRu₂. Previous PES studies of these compounds revealed very similar spectra for the Ce 4*f* electronic states^{8–13}, yet it is expected that such states should be different owing to their differing degrees of hybridization with other valence bands. Our determination of the bulk Ce 4*f* electronic states of these compounds resolves these differences.

In weakly hybridized Ce compounds, where the valence of the Ce ions is close to 3 + (f^1), the Ce 4*f* electrons are localized due to their strong Coulomb interaction. In more strongly hybridized cases, however, the 4*f* electrons form the so-called Kondo singlet with conduction electrons due to the Kondo effect, and behave like heavy fermions at low temperatures³. The strength of hybridization can be represented by the Kondo temperature, T_K ; hybridization is stronger in compounds with higher T_K . A typical heavy fermion system is CeRu₂Si₂, with $T_K \approx 22$ K (refs 14, 15). This material is thought to

be located at a boundary between the localized and itinerant Ce 4*f* states. In more strongly hybridized compounds, on the other hand, the Ce ions are valence-fluctuating between Ce³⁺ (f^1) and Ce⁴⁺ (f^0) states. In such cases, the 4*f* electrons may no longer be localized, and can be regarded as fairly itinerant “4*f*-bands”¹⁶. CeRu₂ is a typical material of the strongly valence-fluctuating 4*f* systems^{17,18}, where T_K is of the order of 1,000 K.

PES probes the photoelectrons produced by incident soft X-ray, or vacuum-ultraviolet, radiation. If valence-band electrons are quite itinerant (corresponding to the weakly correlated case), the valence-band spectra indicate the occupied bands predicted by band-structure calculations. However, spectra of localized Ce 4*f* states are generally inconsistent with the results of band-structure calculations, and are usually explained by the single impurity Anderson model (SIAM)^{19,20} schematically shown in Fig. 1; this represents the localized 4*f* picture. In SIAM, the ground state is a linear combination of the f^0 and $f_{5/2}^1$ states due to the finite hybridization. There can be several 4*f*-photoelectron final states: the f^0 final state which is the so-called poorly screened state, and the well-screened $f_{5/2}^1$ and $f_{7/2}^1$ final states which correspond to a tail of the Kondo peak and its spin-orbit partner. Such peaks can be observed in Ce 4*f* photoemission spectra. The f^0 peak is generally considered to be weak when the hybridization is strong. Furthermore, it has been shown that the tail of the Kondo peak can be observed as an extremely narrow peak in the vicinity of the Fermi energy (E_F) in the 4*f* spectra of rather hybridized Ce compounds^{8,21}. However, the SIAM approach has been the subject of controversy^{11–13}.

It has been difficult to obtain reliable high-resolution Ce 4*f* spectra because the photoionization cross-section, $\sigma(h\nu)$ where $h\nu$ is photon energy, of the Ce 4*f* states is relatively weak in comparison with the σ of other valence-bands at usual $h\nu$ values (such as HeI, HeII etc.)²². Nor does the subtraction method between the two spectra obtained at HeI and HeII provide the correct 4*f* spectrum, because σ of the components other than the 4*f* states is $h\nu$ dependent. So to date, a high-resolution Ce 4*d*-4*f* resonance photoemission (RPES) technique, with energy resolution of 50–100 meV ($h\nu \approx 120$ eV), has been employed with synchrotron

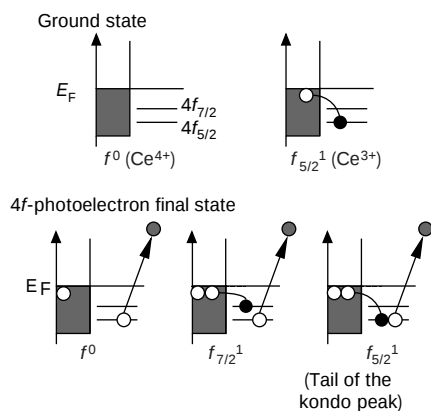


Figure 1 Schematic picture of the single impurity Anderson model (SIAM). Filled and open circles represent the occupied electron and the hole, respectively, while shaded circles stand for the photoelectron. E_F denotes the Fermi level.

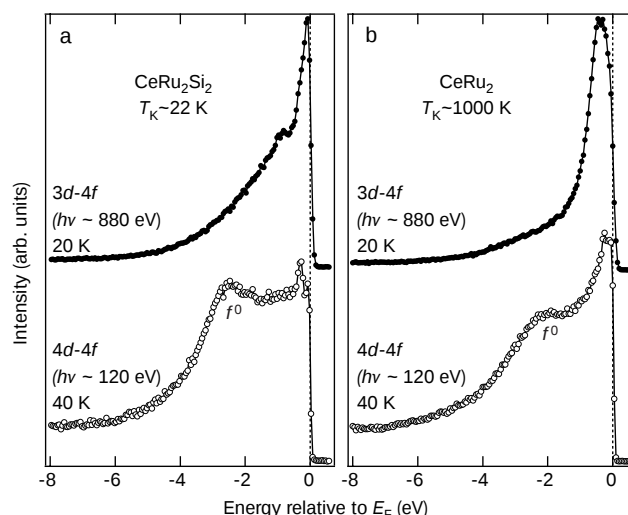


Figure 2 Ce 4*f* spectra obtained from Ce 3*d*-4*f* and 4*d*-4*f* resonance photoemission, with energy resolution of 200 and 80 meV, respectively. **a**, CeRu₂Si₂; **b**, CeRu₂. The Ce 4*f* contributions are estimated by subtracting the off-resonance spectra taken at $h\nu = 875$ (114) eV from the on-resonance ones at $h\nu = 882.6$ (122) eV for the Ce 3*d*-4*f* (4*d*-4*f*) resonance. The relative intensity between the 3*d*-4*f* and 4*d*-4*f* spectra is arbitrary. The zero level of photoemission intensity is shifted for clarity. The surface contribution to the 4*f* spectra is $\sim 55\%$ for the 4*d*-4*f* resonance and 15% for the 3*d*-4*f* resonance.

radiation, because the Ce 4*f* contributions are remarkably enhanced in RPES spectra. However, there is a serious problem for these measurements in that the surface contribution is predominant, and the 4*f* states in the surface are very different from those in the bulk^{9,12}. We note that angle-resolved photoemission spectroscopy (ARPES)—with a changing angle of detection relative to the surface normal of well-defined flat surfaces, and/or surface modification with an isoelectronic non-*f* monolayer²³, are thought to facilitate the deconvolution of the surface components. These methods probe the bulk states indirectly and are not fully applicable to rough surfaces and alloy surfaces. In view of its larger bulk sensitivity, the Ce 3*d*-4*f* RPES is a more direct and promising technique for revealing the bulk Ce 4*f* states: however, the energy resolution of most of the recent such studies has been insufficient (~ 0.7 eV) to observe the tail of the Kondo peak^{9,12}. In order to avoid the above problems, we have constructed a high-resolution soft X-ray (500–1,500 eV) photoemission system; this system consists of a GAMMADATA-SCIENTIA SES-200 spectrometer combined with a varied line-spacing plane grating monochromator on the high-brilliance twin-helical undulator beam line BL25SU in the Spring-8 synchrotron⁷ facility.

The high-resolution Ce 3*d*-4*f* RPES was performed at BL25SU on single-crystal samples, which were prepared by the Czochralski method and annealed under ultrahigh vacuum of 10^{-10} torr. The overall energy resolution was ~ 100 meV at $h\nu \approx 880$ eV. The energy calibration was carefully made to within ± 5 meV by the Fermi cut-off of Au evaporated onto the samples. The samples were cooled to 20 K by a closed-cycle He refrigerator. The results were compared with the high-resolution Ce 4*d*-4*f* RPES spectra measured at BL-3B of the Photon Factory (High Energy Accelerator Research Organization, in Tsukuba, Japan) with the overall resolution of 50 meV. Clean surfaces were obtained either by scraping or by

cleavage *in situ* at low temperatures; surface cleanliness was confirmed before and after the measurements, and the base pressure was about 4×10^{-10} torr. The reproducibility of the spectra was checked several times.

In Fig. 2 we show the Ce 4*f* spectra of CeRu₂Si₂ and CeRu₂ over a wide energy range that includes the Ce 3*d*-4*f* and 4*d*-4*f* resonance (hereafter 3*d*-4*f* and 4*d*-4*f* spectra). In the 3*d*-4*f* spectrum of CeRu₂Si₂, there is a sharp peak near E_F and a broad tail ranging from -1 to -5 eV. According to SIAM, the former corresponds to the contribution of both the tail of the Kondo peak ($f_{5/2}^1$) and its spin-orbit partner ($f_{7/2}^1$). The latter is ascribed to the f^0 final state. In the 4*d*-4*f* spectrum on the other hand, the f^0 final state located at -2.5 eV is prominent, reflecting the localized character of the surface 4*f* states. In the 3*d*-4*f* spectrum of CeRu₂, the f^0 broad tail is greatly suppressed compared to CeRu₂Si₂, indicating that the bulk Ce 4*f* states are considerably hybridized. But the surface f^0 state of CeRu₂ is seen as a hump near -2 eV in the 4*d*-4*f* spectrum, as in CeRu₂Si₂. These differences reveal that the 4*d*-4*f* spectra mainly reflect the surface 4*f* electronic states.

The detailed high-resolution spectra near E_F are shown in Fig. 3 for the 3*d*-4*f* and 4*d*-4*f* resonances. We note that the 4*d*-4*f* spectral line-shapes of CeRu₂Si₂ and CeRu₂ are qualitatively similar, despite the T_K values of these compounds differing by orders of magnitude. There are two comparable structures, located just below E_F and near -0.3 eV. However, the 3*d*-4*f* spectra of these compounds differ greatly. For CeRu₂Si₂, there is a prominent peak in the vicinity of E_F and a weak shoulder near -0.3 eV. These two structures originate from the bulk $f_{5/2}^1$ and $f_{7/2}^1$ final states; in other words, from the tail of the Kondo peak and its spin-orbit partner, as predicted from SIAM. The observation of such a strong tail of the bulk Kondo peak has been made possible by the bulk-sensitive 3*d*-4*f* RPES technique with an unprecedentedly high resolution of 100 meV. The two peaks in the 4*d*-4*f* spectrum of CeRu₂Si₂ can also be assigned in the same manner, but to the surface $f_{5/2}^1$ and $f_{7/2}^1$ final states resulting from the weaker hybridization effect in the surface.

We can also assign the two peaks in the 4*d*-4*f* spectrum of CeRu₂ in Fig. 3 as the surface $f_{5/2}^1$ and $f_{7/2}^1$ states. The 3*d*-4*f* line-shape of CeRu₂ is, however, rather surprising as no structure is seen except for a broad feature centred at -0.5 eV. One would have expected that the tail of the Kondo peak should be much stronger in the bulk-sensitive 3*d*-4*f* spectrum for a very strongly hybridized system like CeRu₂, as predicted from SIAM. The experimental results that the 3*d*-4*f* spectrum shows a rather conventional Fermi cut-off and a broad peak at -0.5 eV cannot be deduced from SIAM. Therefore we conclude that the bulk 4*f* spectral line-shape in CeRu₂ represents itinerant 4*f*-band character due to a very strong hybridization effect, beyond the framework of SIAM. Such spectral behaviour, which has not been revealed by previous 4*d*-4*f* resonance photoemission studies, reflects the real T_K and bulk properties.

We believe that precise measurements of temperature dependence, and ARPES with very high angular resolution, will reveal further details of correlated 4*f* electronic states. The technique that we report here could also be applied to such systems as strongly correlated transition metal compounds. □

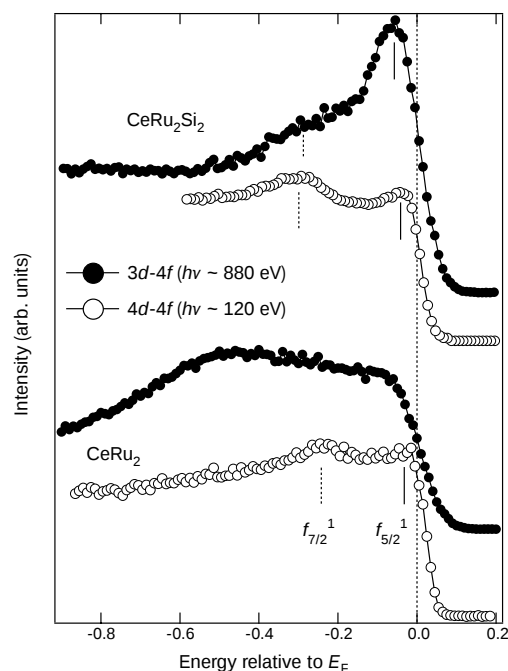


Figure 3 High-resolution Ce 4*f* spectra in the 3*d*-4*f* (resolution, 100 meV) and 4*d*-4*f* (resolution, 50 meV) resonances of CeRu₂Si₂ and CeRu₂. The relative spectral intensity between the 3*d*-4*f* and 4*d*-4*f* resonances and between the compounds is arbitrary. Here the raw spectra taken at $h\nu = 882.6$ eV are displayed for the Ce 3*d*-4*f* spectra since the Ce 4*f* resonance enhancement is predominant. The Ce 4*f* contributions to the 4*d*-4*f* resonance are obtained by subtracting the off-resonance spectra from the one-resonance ones.

Received 16 September; accepted 2 December 1999.

1. Bednorz, J. G. & Müller, K. A. Possible high T_c superconductivity in the Ba-La-Cu-O system. *Z. Phys. B* **64**, 189–193 (1986).
2. Sato, N. *et al.* Magnetic properties of single crystals of Ce₃La_{1-x}B₂. *J. Phys. Soc. Jpn* **53**, 3967–3979 (1986).
3. Onuki, Y. & Komatsubara, T. Heavy fermion state CeCu₆. *J. Magn. Magn. Mater.* **63** & **64**, 281–288 (1987).
4. Nakamura, M. *et al.* Opening of a correlation-induced band gap in NiS. *Phys. Rev. Lett.* **73**, 2891–2894 (1994).
5. Kuwahara, H. *et al.* A first-order phase transition induced by a magnetic field. *Science* **270**, 961–963 (1995).
6. Tanuma, S., Powell, C. J. & Penn, D. R. Proposed formula for electron inelastic mean free paths based on calculations for 31 materials. *Surf. Sci.* **192**, L849–L857 (1987).
7. Saitoh, Y. *et al.* Twin helical undulator beamline for soft x-ray spectroscopy at SPring-8. *J. Synchrotron Rad.* **5**, 542–544 (1998).

8. Patthey, F. *et al.* High-resolution photoemission study of the low-energy excitations in 4f-electron systems. *Phys. Rev. B* **42**, 8864–8881 (1990).
9. Weschke, E. *et al.* Surface and bulk electronic structure of Ce metal studied by high-resolution resonant photoemission. *Phys. Rev. B* **44**, 8304–8307 (1991).
10. Kaindl, G. *et al.* Status and perspectives of high-resolution spectroscopy in the soft x-ray range. *Rev. Sci. Instrum.* **63**, 1234–1240 (1992).
11. Joyce, J. J. *et al.* Temperature-invariant photoelectron spectra in cerium heavy-fermion compounds: Inconsistencies with the Kondo model. *Phys. Rev. Lett.* **68**, 236–239 (1992).
12. Duò, L. *et al.* The Ce 4f surface shift: A test for the Anderson-impurity Hamiltonian. *Phys. Rev. B* **54**, R17363–R17366 (1996).
13. Garnier, M. *et al.* Applicability of the single impurity model to photoemission spectroscopy of heavy fermion Ce compounds. *Phys. Rev. Lett.* **78**, 4127–4130 (1997).
14. Amato, A. *et al.* Thermopower and magneto-thermopower of CeRu₂Si₂ single crystals. *J. Magn. Magn. Mater.* **76 & 77**, 263–264 (1988).
15. Aoki, H., Uji, S., Albessard, A. K. & Ônuki, Y. Observation of heavy electrons in CeRu₂Si₂ via the dHvA effect. *J. Phys. Soc. Jpn* **61**, 3457–3461 (1992).
16. Hedo, M. *et al.* Magnetoresistance and de Haas-van Alphen oscillation in normal and superconducting CeRu₂. *Phil. Mag. B* **77**, 975–1000 (1998).
17. Yang, S.-H. *et al.* High-resolution photoemission study of CeRu₂: The dual character of 4f electrons. *Phys. Rev. B* **53**, R11946–R11948 (1996).
18. Kang, J.-S. *et al.* Photoemission study of an f-electron superconductor: CeRu₂. *Phys. Rev. B* **60**, 5348–5353 (1999).
19. Gunnarsson, O. & Schönhammer, K. Photoemission from Ce compounds: Exact model calculation in the limit of large degeneracy. *Phys. Rev. Lett.* **50**, 604–607 (1983).
20. Allen, J. W. *et al.* Electronic structure of cerium and light rare-earth intermetallics. *Adv. Phys.* **35**, 275–316 (1986).
21. Bickers, N. E., Cox, D. L. & Wilkins, J. W. Thermodynamic, transport and excitation properties of Ce impurities in a model metal: Kondo resonance and universality in the mixed-valent regime. *Phys. Rev. Lett.* **54**, 230–233 (1985).
22. Yeh, J. J. & Lindau, I. Atomic subshell photoionization cross sections and asymmetry parameters $1 \leq Z \leq 103$. *At. Data Nucl. Data Tables* **32**, 1–155 (1985).
23. Witkowski, N. *et al.* Resonant photoemission study of the 4f spectral function of cerium in Ce/Fe(100) interfaces. *Phys. Rev. B* **56**, 12054–12057 (1997).

Acknowledgements

We thank S. Ueda, M. Kotsugi, K. Kadono and R.-J. Jung, and the staffs of SPring-8 and PF, especially T. Matsushita, T. Nakatani, T. Muro, K. Mamiya and A. Kakizaki. We also thank S. Imada for discussions. This work was supported by a Grant-in-Aid for COE Research from the Ministry of Education, Science, Sports and Culture of Japan. The work was performed under the approvals of the PF PAC and of the Japan Synchrotron Radiation Research Institute.

Correspondence and requests for materials should be addressed to A.S.
(e-mail: sekiyama@mp.es.osaka-u.ac.jp).

Dynamic instabilities and memory effects in vortex matter

Y. Paltiel*, E. Zeldov*†, Y. N. Myasoedov*, H. Shtrikman*,
S. Bhattacharya‡§, M. J. Higgins‡, Z. L. Xiao||, E. Y. Andreil||,
P. L. Gammel† & D. J. Bishop†

* Department of Condensed Matter Physics, The Weizmann Institute of Science, Rehovot 76100, Israel

† Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

‡ NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540, USA

§ Tata Institute of Fundamental Research, Mumbai-400005, India

|| Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08855, USA

The magnetic flux line lattice in type II superconductors serves as a useful system in which to study condensed matter flow, as its dynamic properties are tunable. Recent studies have shown a number of puzzling phenomena associated with vortex motion, including: low-frequency noise^{1–5} and slow voltage oscillations^{3,6}; a history-dependent dynamic response^{7–12}, and memory of the direction, amplitude duration and frequency of the previously applied current^{13,14}; high vortex mobility for alternating current, but no apparent vortex motion for direct currents^{13,15,16}; and strong suppression of an a.c. response by small d.c. bias¹³. Taken

together, these phenomena are incompatible with current understanding of vortex dynamics. Here we report a generic mechanism that accounts for these observations. Our model, which is derived from investigations of the current distribution across single crystals of NbSe₂, is based on a competition between the injection of a disordered vortex phase at the sample edges, and the dynamic annealing of this metastable disorder by the transport current. For an alternating current, only narrow regions near the edges are in the disordered phase, while for d.c. bias, most of the sample is in the disordered phase—preventing vortex motion because of more efficient pinning. The resulting spatial dependence of the disordered vortex system serves as an active memory of the previous history.

In conventional superconductors like NbSe₂, the anomalous phenomena are found^{1,2,7–11,13–15} in the vicinity of the ‘peak effect’ where the critical current I_c increases sharply below the upper critical field H_{c2} (Fig. 1a). The peak effect marks a structural transformation^{1,2,7–11,17} of the vortex lattice: below the peak region an ordered phase (OP) is present, which is dominated by the elastic energy of the lattice and is, therefore, weakly pinned. On approaching the peak region, however, the increased softening of the lattice causes a transition into a disordered vortex phase (DP), which accommodates better to the pinning landscape, resulting in a sharp increase in I_c . In high-temperature superconductors like Bi₂Sr₂CaCu₂O₈, this situation is equivalent to the second magnetization peak¹⁸, where the ordered Bragg-glass phase is believed to transform into a disordered solid^{19–21}. Figure 1a shows the I_c measured at various frequencies. On the high-temperature side of the peak effect, I_c is frequency independent; in this region the disordered vortex phase is thermodynamically stable. In contrast, on the low-temperature side, a significant frequency dependence is observed^{13,15}; in this region, all the unusual vortex response phenomena appear^{1,2,7–11,13–15}. As described below, a dynamic coexistence of the ordered vortex phase and a metastable disordered vortex phase is established in this region in the presence of an applied current. We first outline the proposed mechanism, and then present the experimental evidence.

The first ingredient of the proposed model is the observation that in NbSe₂ the disordered vortex phase can be readily supercooled to below the peak effect by field cooling, where it remains metastable, as the thermal fluctuations are negligible^{8–11}. This supercooled disordered phase is pinned more strongly and displays a significantly larger critical current density (J_c^{dis}) than that of the stable ordered phase (J_c^{ord}). An externally applied current in excess of J_c^{ord} serves as an effective temperature and ‘anneals’ the metastable disordered vortex phase, as observed by transport⁸, magnetic response¹⁰, decoration²², and small-angle neutron scattering (SANS) experiments²³ on NbSe₂.

The second ingredient of the model is the presence of substantial surface barriers²⁴, as observed recently²⁵ in NbSe₂. Consider a steady-state flow of an ordered vortex phase in the presence of a transport current. In the standard platelet strip geometry, in a perpendicular field B , vortices penetrate from one edge of the sample and leave at the opposite edge. In the absence of a surface barrier, vortex penetration does not require any extra force. As a result, the vortices penetrate close to their proper vortex lattice locations, as dictated by the elastic forces of the lattice. In the presence of a surface barrier, however, a large force is required for vortex penetration and exit, and hence much of the applied current flows at the edges in order to provide the necessary driving force^{25–28}. The surface barrier is known to be very sensitive to surface imperfections. Therefore, the penetrating vortices are injected predominantly at the weakest points of the barrier, thus destroying the local order and forming a metastable disordered vortex phase near the edge, which drifts into the sample with the flow of the entire lattice. (Note that steps on the surface or extended defects in inhomogeneous samples could also act as injection points of the

disordered phase). The applied current, therefore, has two effects: the current that flows at the edges causes ‘contamination’ by injecting a disordered vortex phase, while the current that flows in the bulk acts as an annealing mechanism. The observed dynamic instabilities and memory phenomena arise for the fine balance between these two competing processes.

The annealing process is sensitive to the exact location on the $H - T$ phase diagram. Below the peak effect, the disordered vortex phase is highly unstable and therefore its relaxation time τ_r in the presence of a driving force is very short. As a result, it anneals rapidly over a characteristic ‘healing’ length $L_r = v\tau_r$, where v is the vortex lattice drift velocity. The corresponding profile of the local critical

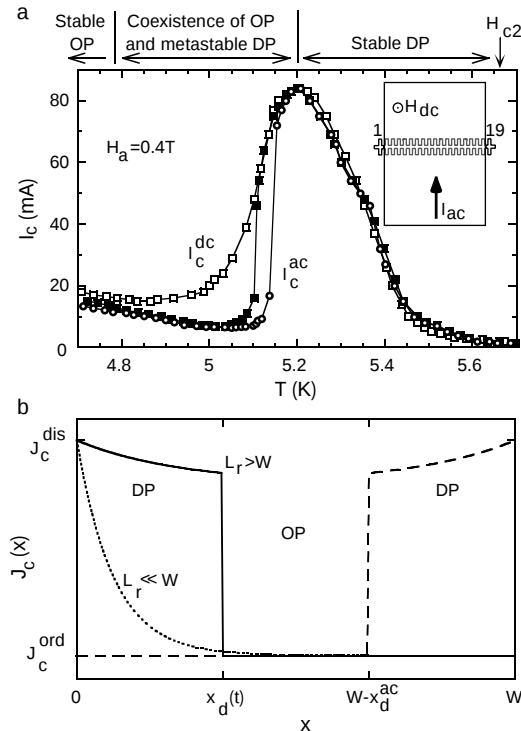


Figure 1 Critical current versus temperature, and the dynamic coexistence of the ordered and disordered phases. **a**, The peak effect in critical current I_c in NbSe₂ crystal A versus temperature, as measured with a d.c. drive, I_c^{dc} (open squares), and a.c. drive, I_c^{ac} , at 181 (filled squares) and 881 Hz (circles). The critical current was determined resistively using a voltage criterion of 1 μV . At low temperatures only the stable ordered vortex phase (OP) is present. On the lower-temperature side of the peak effect, a metastable disordered vortex phase (DP) coexists dynamically with the stable ordered phase, resulting in a frequency-dependent I_c . On the high-temperature side of the peak effect, only the stable disordered phase is present with no anomalous behaviour. Inset, experimental set-up. Several Fe-doped (200 p.p.m.) single crystals of 2H-NbSe₂ were investigated. Here we report data on crystal A of $2.6 \times 0.34 \times 0.05 \text{ mm}^3$ and $T_c = 6.0 \text{ K}$, and crystal B of $2.4 \times 0.29 \times 0.02 \text{ mm}^3$ with $T_c = 6.05 \text{ K}$. Four electrical contacts were attached to the top surface for transport measurements, with the voltage contact separation of $0.6 \pm 0.2 \text{ mm}$. The surface of the crystal was attached to an array of 19 two-dimensional electron gas (2DEG) Hall sensors $10 \times 10 \mu\text{m}^2$ each. The vortex lattice was initially prepared in the ordered phase by zero-field-cooling to a low temperature at which the disordered phase is unstable, and then slowly heated to the desired temperature in the presence of a constant field H_a applied parallel to the c axis. **b**, Schematic plot of the local critical current density $J_c(x)$ across a crystal of width W . J_c^{dis} and J_c^{ord} are the values of the critical current density in the fully disordered phase and in the ordered phase, respectively. For $L_r \ll W$ the disordered phase relaxes rapidly into ordered phase resulting in the dotted $J_c(x)$ in the steady-state d.c. flow. For $L_r > W$, the disordered phase penetrates to a depth $x_d(t)$ following the application of direct current at $t = 0$ (solid curve). For an alternating current at $t = T_{ac}/2$, the disordered phase occupies the left edge to a depth of x_d^{ac} (solid curve), and symmetrically the right edge at $t = T_{ac}$ (dashed curve).

current density $J_c(x)$ should therefore decay from J_c^{dis} to J_c^{ord} over the characteristic length scale L_r , as illustrated by the dotted line in Fig. 1b. Note that L_r and τ_r are generally current dependent, and decrease dramatically at elevated currents⁸. On the other hand, near the peak effect the free energies of the disordered and the ordered phases are comparable, and therefore the ‘life-times’ of the disordered phase τ_r and the corresponding L_r are very large. As a result, the front of the disordered phase, given by $x_d(t)$, progressively penetrates into the bulk as shown by the solid line in Fig. 1b, until the entire sample is contaminated. In this situation the experimental, steady-state d.c. critical current I_c^{dc} does not reflect an equilibrium property, but rather a dynamic coexistence of two phases. It is given by $I_c^{\text{dc}} = d \int_0^W J_c(x) dx \approx dL_r J_c^{\text{dis}} + d(W - L_r) J_c^{\text{ord}}$ for $L_r < W$, and $I_c^{\text{dc}} \approx I_c^{\text{dis}} = dW J_c^{\text{dis}}$ for $L_r \gg W$, where d and W are the thickness and width of the sample, respectively (neglecting, for simplicity, the surface barrier edge currents, and assuming, for example, an exponential decay of $J_c(x)$ in Fig. 1b).

The a.c. response of the system should be distinctly different, because the contamination process occurs only near the edges, where the disordered lattice periodically penetrates and leaves the sample. For a square wave I_{ac} of period $T_{ac} = 1/f$, by the end of the positive half cycle the disordered phase occupies the left edge to a depth of x_d^{ac} , as illustrated by the solid curve in Fig. 1b. During the negative half cycle the disordered phase on the left leaves the sample, while a disordered phase on the right edge penetrates, until at $t = T_{ac}$ a mirror-image profile is obtained, as shown by the dashed line. Assuming $x_d^{\text{ac}} < W < L_r$, the effective I_c observed by a.c. transport measurement is given by $I_c^{\text{ac}} \approx d x_d^{\text{ac}} J_c^{\text{dis}} + d(W - x_d^{\text{ac}}) J_c^{\text{ord}}$. Thus, an alternating current necessarily contaminates the sample less than a direct current of the same amplitude, and therefore I_c^{ac} is always $\leq I_c^{\text{dc}}$, as seen in Fig. 1a. In addition, as x_d^{ac} decreases with frequency, I_c^{ac} should decrease with f , explaining the frequency dependence of I_c^{ac} in Fig. 1a. Furthermore, and most importantly,

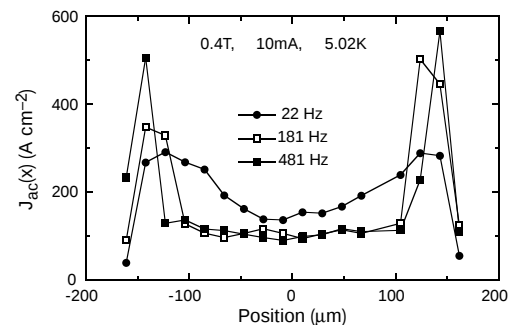


Figure 2 Current density profiles $J_{ac}(x)$ in crystal B. These are obtained by inversion of the self-induced field measured by the Hall sensors. Shown are results for three frequencies f , 22 Hz (filled circles), 181 Hz (open squares), and 481 Hz (filled squares). A square wave alternating current I_{ac} was applied, and the corresponding self-induced magnetic field $B_{ac}(x)$ across the crystal was measured by the Hall sensors using a lock-in amplifier (see Fig. 1 inset). By using the Biot-Savart law, $B_{ac}(x)$ was directly inverted²⁵ to give the current-density profiles $J_{ac}(x)$. The width x_d^{ac} of the highly pinned disordered phase near the edges grows with decreasing frequency, as expected. The measured $J_{ac}(x)$ is the magnitude of the local current density averaged over the a.c. cycle period. As a result, $J_{ac}(x)$ reflects a time-averaged superposition of solid and dashed $J_c(x, t)$ profiles in Fig. 1b, which are present separately during the positive and negative half-cycles. Close to the edges the high J is present most of the time, while close to x_d^{ac} it is present only a small fraction of the a.c. period as the disordered phase moves in and out of the sample. Therefore, the time-averaged $J_{ac}(x)$ decreases smoothly from the edge to x_d^{ac} . A more detailed analysis shows that the first-harmonic measurement by the lock-in amplifier Fourier-transforms the sharp instantaneous $J_c(x, t)$ of Fig. 1b into the observed smooth $J_{ac}(x) = J_c^{\text{ord}} + (J_c^{\text{dis}} - J_c^{\text{ord}})(1 + \cos(\pi x/x_d^{\text{ac}}))/2 + (J_c^{\text{dis}} - J_c^{\text{ord}})(1 - \cos(\pi(x + x_d^{\text{ac}} - W)/x_d^{\text{ac}}))/2$. The second and third terms hold at $0 \leq x \leq x_d^{\text{ac}}$ and $W - x_d^{\text{ac}} \leq x \leq W$, respectively, and are zero otherwise.

at sufficiently high frequency I_c^c should approach the true I_c of the stable phase. The steep increase of the 881-Hz I_{ac} data in Fig. 1a (open circles) therefore indicates that the ordered vortex phase transforms sharply into the disordered phase at the peak effect. In contrast, the smooth behaviour of I_c^d reflects rather the dynamic coexistence of the two phases, in which L_r gradually increases and diverges on approaching the peak effect from below. From Fig. 1a we can evaluate L_r and τ_r . For example, at $T = 5.1$ K, $I_c^d \approx 50$ mA is about half way between $I_c^{ord} \approx 5$ mA and the extrapolated $I_c^{dis} \approx 100$ mA, which means that $L_r \approx 0.5W = 170$ μ m. The I_c^d was measured at a voltage criterion of 1 μ V, which translates into vortex velocity of $v \approx 4 \times 10^{-3}$ m s $^{-1}$, and hence $\tau_r = L_r/v \approx 4 \times 10^{-2}$ s. This value is well within the range of the relaxation times measured previously⁸ by applying a current step to the field-cooled metastable disordered vortex phase.

We now provide a direct experimental manifestation of the central aspect of the model, which is the spatial dependence of the disordered system and $J_c(x)$, and of the transport current distribution that traces this $J_c(x)$ (see Fig. 1b). We have used Hall sensor arrays to measure the a.c. transport current self-induced field^{25–27} $B_{ac}(x)$, which is then directly inverted to give the current density distribution $J_{ac}(x)$ using the Biot–Savart law, as described previously²⁵ (see Fig. 1a inset). Figure 2 shows the corresponding current density profiles $J_{ac}(x)$ measured at different frequencies. At high f , the disordered vortex phase with the enhanced J_c is present only in narrow regions near the edges (481-Hz data). As the frequency is reduced, x_d^{ac} grows and correspondingly the enhanced $J_{ac}(x)$ flows in wider regions near the edges. Note that our measurement procedure (see Fig. 2) provides the time-averaged local amplitude of $J_{ac}(x)$, which is much smoother than the sharp instantaneous profiles in Fig. 1b.

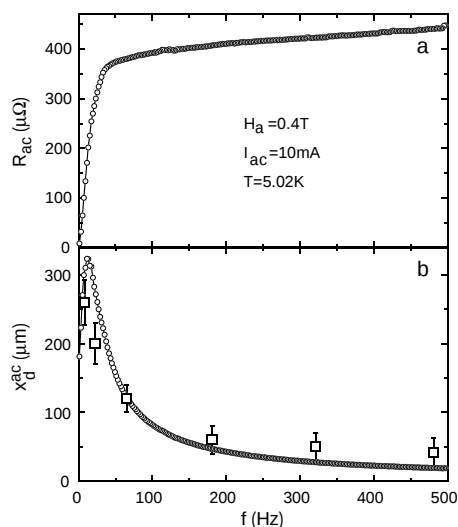


Figure 3 Frequency dependence of the resistance $R_{ac}(f)$ and of the width of the disordered regions x_d^{ac} in crystal B. **a**, At high frequencies R_{ac} is large, as most of the sample is in the weakly pinned ordered phase. As the frequency is decreased the disordered regions increase and R_{ac} drops sharply when x_d^{ac} reaches values close to the sample width. In the limit of zero f the resistance is zero as the applied current is lower than I_c^c . **b**, The corresponding width of the disordered phase near the edges x_d^{ac} derived from the $R_{ac}(f)$ data (circles) and from the $J_{ac}(x)$ current profiles of Fig. 2 (squares). As expected, x_d^{ac} increases monotonically with decreasing frequency. The decrease of the x_d^{ac} data (circles) at very low frequencies is an artefact. At such frequencies, the instantaneous vortex motion is present mainly at the onset of the square wave I_{ac} pulses, and decays towards zero during the pulse¹³. In this situation the first-harmonic R_{ac} measurement by the lock-in amplifier underestimates the integrated vortex displacement. In addition to the frequency dependence, x_d^{ac} also changes significantly when the amplitude of the a.c. current I_{ac} is varied. Here $I_{ac} = 10$ mA was chosen, such that x_d^{ac} becomes comparable to the sample width at low frequencies.

We confirm the above finding independently by measuring the corresponding a.c. resistance of the sample $R_{ac}(f)$ (Fig. 3a). At high frequencies, most of the sample is in the low pinning ordered phase and therefore R is large. As f is decreased, progressively wider regions near the edges become contaminated with the more strongly pinned disordered phase, and thus $R_{ac}(f)$ decreases. If the applied I_{ac} is larger than I_c^c , a finite R will be measured at all frequencies; but if $I_c^c \leq I_{ac} \leq I_c^d$ (see Fig. 1a), the measured R will vanish as $f \rightarrow 0$, as observed in Fig. 3a. This explains the surprising phenomenon of finite vortex response to a.c. current, while for d.c. drive the vortex motion is absent^{13,15,16}. One can directly calculate the width of the disordered regions from $R_{ac}(f)$ by noting that x_d^{ac} equals the distance the entire lattice is displaced during half an a.c. period; $x_d^{ac}(f) = v/2f = R(f)I_{ac}/2fLB$, where L is the voltage contact separation. The open circles in Fig. 3b show x_d^{ac} obtained from $R_{ac}(f)$, while the open squares show the x_d^{ac} derived directly from the $J_{ac}(x)$ profiles of Fig. 2. The good correspondence between the two independent evaluations of x_d^{ac} demonstrates the self-consistency of the model.

Next we address the extreme sensitivity of the a.c. response to a small d.c. bias¹³, as shown in Fig. 4a, where R_{ac} is presented as a function of a superposed I_{dc} . A d.c. bias of only 10% to 20% of I_{ac} suppresses R_{ac} by orders of magnitude. This behaviour is a natural consequence of the described mechanism, because the d.c. bias contaminates the sample very similarly to the pure d.c. case, except that L_r is now renormalized as follows. For $I_{dc} \ll I_{ac}$, the vortices move back and forth during the a.c. cycle, with a forward displacement being larger by about $2I_{dc}/I_{ac}$. Therefore, a vortex that enters through the sample edge and reaches a position accumulates a much longer total displacement path of xI_{ac}/I_{dc} . Because the annealing process of the disordered vortex phase depends on the total

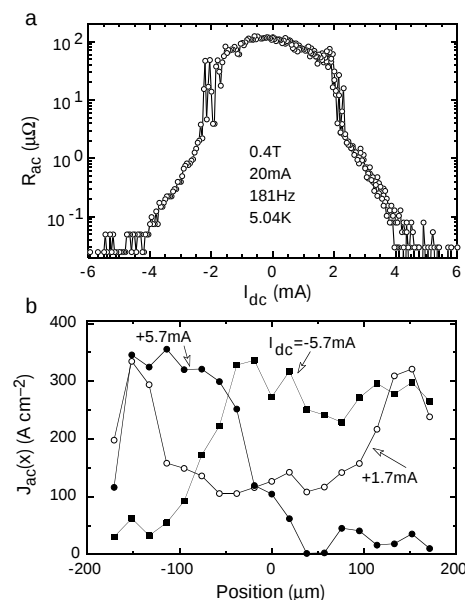


Figure 4 Measured resistance as a function of d.c. bias, and the corresponding distribution of the alternating current in crystal A. **a**, At low d.c. bias, $I_{dc} \leq 2$ mA, only the regions near the edges are contaminated by the disordered vortex phase as L_r^{eff} is very short, resulting in only a moderate decrease of the measured resistance R_{ac} . ($I_{ac} = 20$ mA for all R_{ac} measurements shown here.) For $I_{dc} \geq 2$ mA, the contamination becomes substantial, and the significant decrease of R_{ac} is accompanied by enhanced noise due to the local instabilities during the competing contamination and annealing processes. At still larger d.c. bias, most of the sample becomes contaminated and R_{ac} drops below our noise level. The corresponding $J_{ac}(x)$ profiles in **b** show that for positive $I_{dc} = +5.7$ mA (filled circles) a substantial part of the sample is contaminated from the left edge where the vortices enter the crystal, and similarly from the right edge for negative bias $I_{dc} = -5.7$ mA (filled squares).

displacement regardless of the direction, the lattice at this location is thus annealed substantially, as if the effective L_r is reduced to $L_r^{\text{eff}} \approx L_r I_{dc}/I_{ac}$. Thus, at very small biases, the disordered phase is present only within x_d^{ad} from the edges, as in the absence of a bias, where the disordered vortices leave and re-penetrate every cycle. Vortices that drift deeper into the bulk under the influence of I_{dc} are practically fully annealed because of the very short L_r^{eff} . As a result, the initial decrease of R_{ac} up to $I_{dc} \approx 2$ mA (see Fig. 4a) is relatively small. The corresponding $J_{ac}(x)$ in Fig. 4b at $I_{dc} = 1.7$ mA shows narrow contaminated regions near the edges, very similar to the zero-bias case in Fig. 2. However, as I_{dc} is increased, L_r^{eff} grows and the bulk of the sample becomes contaminated by the penetrating disordered vortex phase, leading to a marked drop of R_{ac} . In this situation, $J_{ac}(x)$ at $I_{dc} = 5.7$ mA shows a wide region of disordered phase at the left edge. When I_{dc} is inverted to -5.7 mA, a similar situation is observed, but now the vortices—and hence the disordered phase—penetrate from the right edge, as expected.

The revealed mechanism readily explains a wide range of additional reported phenomena. (1) The history of the previously applied current is encoded in the spatial profile of the lattice disorder, which is preserved after the current is switched off owing to negligible thermal relaxation. On reapplying the current, the vortex system will display a memory of all the parameters of the previously applied current, including its direction, duration, amplitude and frequency, as observed experimentally^{13,14}. (2) Application of a current step $I < I_{dc}^c$ to a sample in the ordered vortex phase results in a transient response which decays to zero, as the disordered phase is able to penetrate only a limited distance. The resulting new I_c of the sample is given by the condition that $I_c = I$, as derived by fast transport measurements^{14,15}. Such transient phenomena would also display characteristic times shorter than, or comparable to, the vortex transit time across the sample, in agreement with observations^{13,14}. (3) The competition between the contamination and annealing processes is expected to result in local instabilities, causing the reported noise enhancement below the peak effect^{1,2} (see also Fig. 4a). (4) Related phenomena should be observed in high-temperature superconductors in the vicinity of the peak effect associated with the melting transition, or near the second magnetization peak, consistent with experiments^{3–6,12,16}. (5) In high-temperature superconductors there is an additional consideration of thermal activation of vortices over the surface barriers, which may explain the reported slow voltage oscillations^{3–6}. If the thermal activation rate is higher than, or comparable to, the driving rate, the slowly injected lattice will be ordered, in contrast to the disordered vortex phase injected at higher drives. Thus, at a given applied current, if the bulk of the sample is in the ordered vortex phase, much of the current flows on the edges, rapidly injecting a disordered vortex phase through the surface barrier. Once the bulk becomes contaminated, the resulting slower vortex motion again causes injection of an ordered phase. This feedback mechanism can explain the voltage oscillations^{3,6} in $\text{YBa}_2\text{Cu}_3\text{O}_7$ and similar narrow-band noise^{4,5} in $\text{YBa}_2\text{Cu}_3\text{O}_7$ and $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ with characteristic frequencies comparable to the inverse transit time. Last, we note that the described phenomena should be absent in the Corbino disk geometry where vortices do not cross the sample edges. Our studies of NbSe_2 in this geometry (Y.P. *et al.*, manuscript in preparation) confirm this prediction. □

Received 5 August; accepted 16 November 1999.

1. Marley, A. C., Higgins, M. J. & Bhattacharya, S. Flux flow noise and dynamical transitions in a flux line lattice. *Phys. Rev. Lett.* **74**, 3029–3032 (1995).
2. Merithew, R. D. *et al.* Persistent metastable states in vortex flow at the peak effect in NbSe_2 . *Phys. Rev. Lett.* **77**, 3197–3199 (1996).
3. Kwok, W. K. *et al.* Dynamic instabilities in the vortex lattice of $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Physica C* **293**, 111–117 (1997).
4. D'Anna, G. *et al.* Vortex-motion-induced voltage noise in $\text{YBa}_2\text{Cu}_3\text{O}_7$ single crystals. *Phys. Rev. Lett.* **75**, 3521–3524 (1995).
5. Tsuboi, T., Hanaguri, T. & Maeda, A. Local density fluctuations of moving vortices in the solid and liquid phases in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **80**, 4550–4553 (1998).

6. Gordeev, S. N. *et al.* Current-induced organization of vortex motion in type-II superconductors. *Nature* **385**, 324–326 (1997).
7. Bhattacharya, S. & Higgins, M. J. Flux-flow fingerprint of disorder: melting versus tearing of a flux-line lattice. *Phys. Rev. B* **52**, 64–67 (1995).
8. Henderson, W. *et al.* Metastability and glassy behavior of a driven flux-line lattice. *Phys. Rev. Lett.* **77**, 2077–2080 (1996).
9. Banerjee, S. S. *et al.* Anomalous peak effect in CeRu_2 and 2H-NbSe_2 : Fracturing of a flux line lattice. *Phys. Rev. B* **58**, 995–999 (1999).
10. Banerjee, S. S. *et al.* Metastability and switching in the vortex state of 2H-NbSe_2 . *Appl. Phys. Lett.* **74**, 126–128 (1999).
11. Wordenweber, R., Kes, P. H. & Tsuei, C. C. Peak and history effect in two-dimensional collective flux pinning. *Phys. Rev. B* **33**, 3172–3180 (1986).
12. Kokkalis, S. *et al.* Onset of plasticity and hardening of the hysteretic response in the vortex system of $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$. *Phys. Rev. Lett.* **82**, 5116–5119 (1999).
13. Henderson, W., Andrei, E. Y. & Higgins, M. J. Plastic motion of a vortex lattice driven by alternating current. *Phys. Rev. Lett.* **81**, 2352–2355 (1998).
14. Xiao, Z. L., Andrei, E. Y. & Higgins, M. J. Flow induced organization and memory of a vortex lattice. *Phys. Rev. Lett.* **83**, 1664–1667 (1999).
15. Andrei, E. Y. *et al.* Current driven organization of magnetic vortices. *J. Phys. IV* **10**, 5–10 (1999).
16. Metlushko, V. *et al.* Driven vortex states and relaxation in single crystal $\text{YBa}_2\text{Cu}_3\text{O}_8$. Preprint cond-mat/9804121 at (<http://xxx.lanl.gov>) (1999).
17. Gammel, P. L. *et al.* Structure and correlation of a flux line lattice in crystalline Nb through peak effect. *Phys. Rev. Lett.* **80**, 833–836 (1998).
18. Khaykovich, B. *et al.* Vortex-lattice phase transitions in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ crystals with different oxygen stoichiometry. *Phys. Rev. Lett.* **76**, 2555–2558 (1996).
19. Giamarchi, T. & Le Doussal, P. Elastic theory of pinned flux lattices. *Phys. Rev. Lett.* **72**, 1530–1533 (1994).
20. Ertas, D. & Nelson, D. R. Irreversibility, mechanical entanglement and thermal melting in superconducting vortex crystal with point impurities. *Physica C* **272**, 79–85 (1996).
21. Vinokur, V. *et al.* Lindemann criterion and vortex-matter phase transitions in high-temperature superconductors. *Physica C* **295**, 209–217 (1998).
22. Pardo, F. *et al.* Topological defects in the flux-line lattice and their relationship to the critical current of a type-II superconductor. *Phys. Rev. Lett.* **78**, 4633–4636 (1996).
23. Yaron, U. *et al.* Structural evidence for a two-step process in the depinning of the superconducting flux-line lattice. *Nature* **376**, 753–755 (1995).
24. Bean, C. P. & Livingston, J. D. Surface barrier in type-II superconductors. *Phys. Rev. Lett.* **12**, 14–16 (1964).
25. Paltiel, Y. *et al.* Surface barrier dominated transport in NbSe_2 . *Phys. Rev. B* **58**, R14763–R14766 (1998).
26. Fuchs, D. T. *et al.* Transport properties governed by surface barriers in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Nature* **391**, 373–376 (1998).
27. Fuchs, D. T. *et al.* Possible new vortex matter phases in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **80**, 4971–4974 (1998).
28. Burlachkov, L., Koshelev, A. E. & Vinokur, V. M. Transport properties of high-temperature superconductors: surface vs. bulk effect. *Phys. Rev. B* **54**, 6750–6757 (1996).

Acknowledgements

We thank P. B. Littlewood for discussions. The work at the Weizmann Institute of Science was supported by the Israel Science Foundation—Centre of Excellence Program, by the US-Israel Binational Science Foundation (BSF), and by the Alhadeff research award. E.Y.A. was supported by the NSF.

Correspondence and requests for materials should be addressed to Y.P. (e-mail: hpalt@wis.weizmann.ac.il).

Singularity dynamics in curvature collapse and jet eruption on a fluid surface

Benjamin W. Zeff*, Benjamin Kleber*, Jay Fineberg† & Daniel P. Lathrop*

* Institute for Plasma Research and Department of Physics, The University of Maryland, College Park, Maryland 20742, USA

† The Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel

Finite-time singularities—local divergences in the amplitude or gradient of a physical observable at a particular time—occur in a diverse range of physical systems. Examples include singularities capable of damaging optical fibres and lasers in nonlinear optical systems¹, and gravitational singularities² associated with black holes. In fluid systems, the formation of finite-time singularities

cause spray and air-bubble entrainment³, processes which influence air–sea interaction on a global scale^{4,5}. Singularities driven by surface tension have been studied in the break-up of pendant drops^{6–9} and liquid sheets^{10–12}. Here we report a theoretical and experimental study of the generation of a singularity by inertial focusing, in which no break-up of the fluid surface occurs. Inertial forces cause a collapse of the surface that leads to jet formation; our analysis, which includes surface tension effects, predicts that the surface profiles should be describable by a single universal exponent. These theoretical predictions correlate closely with our experimental measurements of a collapsing surface singularity. The solution can be generalized to apply to a broad class of singular phenomena.

In our experiment, surface singularities are produced by the collapsing depressions of standing waves, a phenomenon first observed by Longuet-Higgins¹³. A cylindrical tank, partially filled with a fluid, is vertically oscillated with an acceleration of the form

$a(t) = A \sin(\omega t)$, where A is the acceleration amplitude, and ω is the driving frequency. This forcing, known as Faraday excitation, generates subharmonic standing waves on the fluid surface¹⁴. Below a critical standing wave height h_c the fluid surface topology remains smooth and simply connected. Above h_c , the collapsing wave entrains an air bubble beneath the surface and changes its topology from simply to multiply connected. Thus, the critical height represents the threshold of the surface topology change. At this point, where parameter values are just below those values for which droplet entrainment occurs, the resulting inertial collapse creates a singularity on the fluid surface at which the velocity and surface curvature diverge. This singularity focuses the kinetic energy of the fluid along the central axis and produces a narrow, high-speed vertical jet (Fig. 1). Experimentally, we observe the approach to this singularity; as with many such systems, an ultraviolet cutoff exists which prevents a full divergence of the physical variables. In this case, viscous forces¹⁵ become important near in time and space to the singularity and provide one source for this cutoff. In addition, the Rayleigh instability in the jet causes the formation of a droplet at the tip and eventually causes the jet to break into droplets.

In order to simplify nearly intractable nonlinear equations, the fluid is assumed to be irrotational and incompressible, and the air above the fluid is considered to be of zero density. Near the location of the event, the surface tension forces, kinetic energy density, and acceleration are all thought to diverge jointly, and gravity may safely

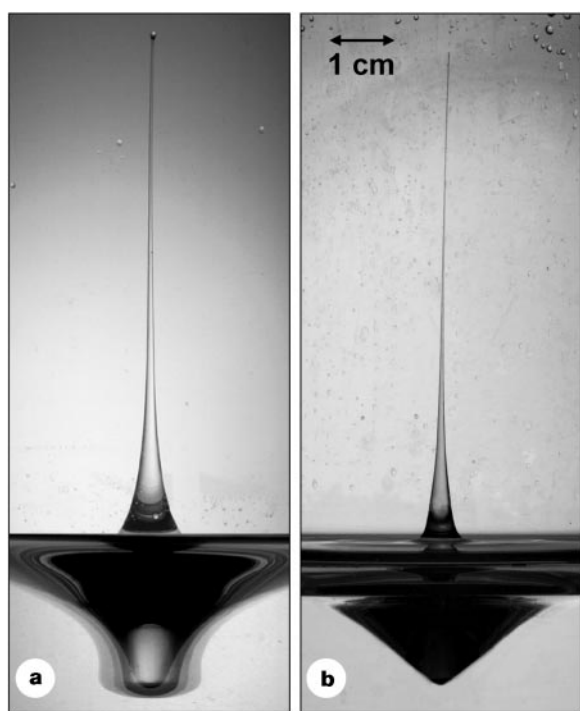


Figure 1 Surface wave collapses and resulting jets. **a**, A composite of two photographs showing the collapse of a surface wave depression and the subsequent upward jet caused by self-focusing of the kinetic energy associated with a near singularity. The singularities are produced in a cylindrical tank with an inner diameter of 12.7 cm filled to a depth of 6.5 cm with a glycerin–water solution with a viscosity of $1.94 \text{ cm}^2 \text{ s}^{-1}$. A sinusoidal signal of frequency 7.84 Hz generated by computer controls a servo circuit driving the linear actuator which holds the tank. Initially, the system is driven at a low amplitude, 2.39 m s^{-2} , which sets up a repeatable periodic standing wave. Once the wave state settles into a periodic state, the driving amplitude is increased suddenly at a zero crossing to 3.23 m s^{-2} (or nearby values to produce different wave states), after which the height of the standing waves increases with each subsequent period. Once the standing wave is sufficiently tall, it produces a deep depression which collapses to a singularity and jet. Images of the developing singularity and jet are captured with a $2,000 \times 3,000$ pixel digital camera and a five-flash array which are triggered via computer signal and a digital delay generator. With this system, five exposures separated in time by as little as 250 μs can be captured on a single image. Here, the first photograph is a four-exposure image and shows four time steps of the collapse which correspond to 8 ms, 4 ms, 2 ms, and 1 ms before the singularity. The second photograph shows the jet 33 ms after the singularity. **b**, A second composite, here at a viscosity of $0.26 \text{ cm}^2 \text{ s}^{-1}$. The driving frequency and amplitude are 8.00 Hz and 4.64 m s^{-2} , respectively. The conical shape of the surface at the time of the singularity is shown with a resultant jet. In **a**, viscous effects provide an ultraviolet cutoff which smoothes the asymptotic cone.

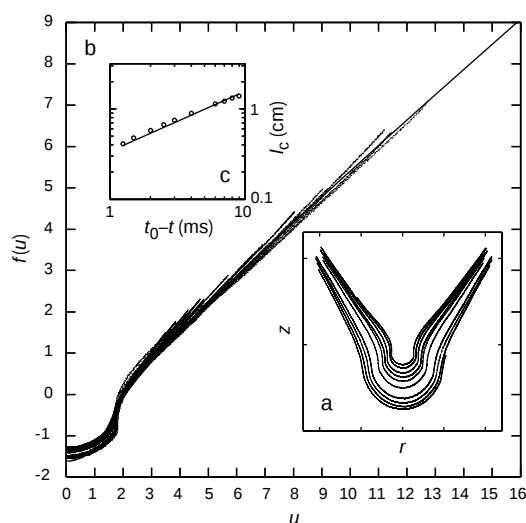


Figure 2 Comparison of theoretical and experimental surface profiles for $0.26 \text{ cm}^2 \text{ s}^{-1}$ glycerin–water solution. **a**, The surface profile of a collapsing depression at ten times before the singularity: 9 ms, 8 ms, 7 ms, 6 ms, 4 ms, 3 ms, 2.5 ms, 2 ms, 1.5 ms, and 1.25 ms. Curves closer to the central axis occur later in time. The base of depression accelerates upward and reaches the origin at the time t_0 . The standing wave which produces this depression is subcritical in height, so it collapses to a near singularity. The surface profiles are converted from digital images to coordinates and corrected for the refraction caused by the cylindrical tank. The height of the waves and qualitative information about the events are determined from images captured by a 250 frames/second video camera. **b**, The theoretical form of the surface profile, calculated numerically as described in the text for $We = 4,000$, is shown as a solid line. The asymptotic cone angle is $\theta = 60^\circ$. The points are the ten surface profiles from **a** scaled according to the similarity solution. The profiles collapse into one self-similar form which agrees well with theory. Because of a viscous ultraviolet cutoff, overhang begins to develop in the surface profiles at late times ($<1 \text{ ms}$). This overhang is not present in the numerical form. **c**, A log–log plot of l_c , the mean distance from the origin to all points (with $f(u) < 0$, where $f(u) = h(r, t)(t_0 - t)^{-2/3}$ is the scaled height of the fluid surface) along a given profile, and $(t_0 - t)$ for each of the curves in **a** and **b**. The solid line represents the predicted two-thirds scaling law.

be neglected. An important result of these assumptions is that the physical forcing of the system no longer enters into the local equations of motion. Hence, the same equations describe the curvature collapse and singularities in many other systems with free surfaces, independent of forcing, including spherical and cylindrical cavitation bubble collapse, droplet impact rebounds and, by considering the gravity-dominated case, wave impact and jets on a shear wall. We show that the inclusion of surface tension not only expands on previous work^{6,17}, but is also responsible for the selection of a universal solution describing jet formation in collapsing cavities. The system can be described by three equations, one for the bulk and two at the air–fluid interface: the Laplace equation representing incompressibility, the kinematic equation for the fluid surface, and the Bernoulli equation at the fluid surface. These three equations in cylindrical coordinates are, respectively:

$$\nabla^2 \phi = 0 \quad (1)$$

$$\frac{\partial h}{\partial t} + \frac{\partial h}{\partial r} \frac{\partial \phi}{\partial r} = \frac{\partial \phi}{\partial z} \quad (2)$$

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}(\nabla \phi)^2 + \frac{\sigma}{\rho} \left(\frac{1}{R_1} + \frac{1}{R_2} \right) = 0 \quad (3)$$

Here ϕ is the velocity potential ($\mathbf{v} = \nabla \phi$), z is the axial coordinate ($z = h$ at the free surface), r is the radial distance, σ and ρ are the surface tension and density of the fluid, respectively, and R_1 and R_2 are the principal radii of curvature of the surface.

There is no formal mathematical framework for attacking such a system of nonlinear partial differential equations. However, in past analysis of singularities^{6,7,18,19}, a similarity method has proven an effective tool. We postulate a power-law scaling in times close to the singularity:

$$h(r, t) = (t_0 - t)^\alpha f[r(t_0 - t)^\beta] \quad (4)$$

$$\phi(r, z, t) = (t_0 - t)^\gamma g[r(t_0 - t)^\beta, z(t_0 - t)^{-\alpha}], \quad (5)$$

in which t_0 is the time of the singularity. The variables r and z have been scaled by the characteristic length and velocity scales l and v , and t has been scaled by a characteristic time $t = l/v$. This scaling yields a single non-dimensional parameter, the Weber number $We = \rho v^2 l / \sigma$ that governs the dynamics of the collapse. Physically, We is the ratio of the kinetic energy of the fluid motion to the surface energy that contains the fluid. We estimate the Weber number based

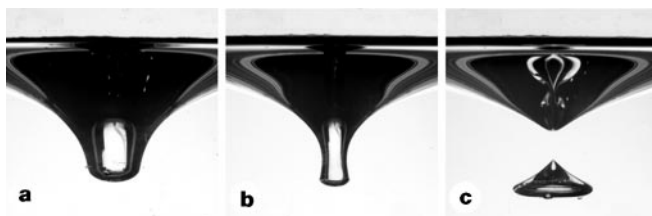


Figure 3 Surface depressions produced by surface waves of three different heights in $1.94 \text{ cm}^2 \text{ s}^{-1}$ fluid. **a**, The depression produced by a collapsing wave of subcritical height, $h < h_c$. The surface topology remains smooth and simply connected. A jet is produced with a tip velocity of 7 m s^{-1} . **b**, The depression produced by a standing wave with $h \approx h_c$. The walls are nearly vertical, and no bubble is entrained. This state produces a very narrow jet with a velocity of 30 m s^{-1} . **c**, The depression produced by a standing wave with $h > h_c$. A large conical bubble is produced as the topology changes from simply to multiply connected. The curvature is singular at the point of bubble pinch-off. The bubble decreases the kinetic energy of this jet, which has a velocity of 6 m s^{-1} . Large bubble oscillations and concomitant emitted sound account for much of the energy loss. Each of these photographs was taken at the same time relative to the phase of the forcing.

on the height of the last smooth wave h and the velocity $v = h\omega/2$ (where $\omega/2$ is the angular frequency of the wave).

For a similarity solution to exist, the time-dependence of all terms in equations (1)–(3) must be the same. Substituting equations (4) and (5) and requiring the time-dependence to be identical in each term, we find that only a single value is allowed for each of the exponents: $\alpha = 2/3$, $\beta = -2/3$, and $\gamma = 1/3$. These exponents were also found by Keller and Miksis¹⁰ in their analysis of a breaking liquid sheet and apply to all cases when only surface tension and inertial forces are dominant. This result differs significantly from the case in which surface tension is neglected^{16,17}, in which, for given conditions, a continuous family of exponents is possible with no well defined selection criterion.

We can infer important characteristics of the singularity from the values of the scaling exponents, including the profile of the fluid surface near in time to the singularity. As $t_0 - t \rightarrow 0$, the scaled radius in the argument of f diverges. Since $\alpha = -\beta$, the function f must be linear in its argument for times near the singularity in order for the height to remain finite. Near the singularity, then, the height of the surface is linear with respect to the radius; that is, $h(r, t) \rightarrow Cr$ and the surface is a cone as $t \rightarrow t_0$ (Fig. 1b). Numerical evidence indicates that only certain values of the cone angle $C = \cot \theta$ (the angle of the surface off the vertical) are allowed. The system is now described by three partial differential equations for f and g in the scaled radius $u = r(t_0 - t)^{-2/3}$ and height $v = z(t_0 - t)^{2/3}$.

$$\nabla^2 g = 0 \quad (6)$$

$$-\frac{2}{3}f + \frac{2}{3}u \frac{\partial f}{\partial u} + \frac{\partial f}{\partial u} \frac{\partial g}{\partial v} = \frac{\partial g}{\partial v} \quad (7)$$

$$-\frac{1}{3}g + \frac{2}{3}u \frac{\partial g}{\partial u} + \frac{2}{3}f \frac{\partial g}{\partial v} + \frac{1}{2} \left(\frac{\partial g}{\partial u} \right)^2 + \frac{1}{2} \left(\frac{\partial g}{\partial v} \right)^2 + \frac{1}{We} \left(\frac{1}{R_1} + \frac{1}{R_2} \right) = 0 \quad (8)$$

To solve equations (7) and (8) numerically and find a form for the surface profile (Fig. 2b), we write equation (6) as a series of known solutions to the Laplace equation with unknown coefficients:

$$g = c_{1/2} r_s^{1/2} P_{1/2}(\cos \theta) + c_{-1} r_s^{-1} + c_{-2} r_s^{-2} P_1(\cos \theta) + \dots \quad (9)$$

in which r_s is the spherical radius from the centre of the collapsing cone, θ is the polar angle, and P_n are Legendre polynomials. Asymptotic analysis of equation (8) requires the leading-order behaviour to be $r_s^{1/2}$. In solving for $f(u)$ numerically, the coefficients in the above series and the angle θ_0 at large radius are assigned values. This form of g is substituted into equation (2) which is then integrated to determine $f(u)$. The coefficients and angle are varied, and $f(u)$ and $g(u, v)$ are recalculated iteratively until equation (3) is satisfied. This numerical form predicts not only the shape of the collapsing surface in our experiment, but due to the lack of a specific forcing in equations (6)–(8), is a general form which describes any axisymmetric surface singularity dominated by inertia and surface tension.

We experimentally observe the approach to the singularity in the curvature collapse of the surface profile of the fluid surface. Figure 3 shows three collapsing surface depressions in a $1.94 \text{ cm}^2 \text{ s}^{-1}$ fluid in the tank, one subcritical, one nearly critical, and one beyond the critical height. The collapse of all three of these surfaces would produce a fluid jet. Figure 3b shows the collapse of the depression near the critical point which produces an energetic singularity. The walls of the depression collapse radially inward while the bottom of the depression accelerates upward toward an origin. At the point of singularity, the velocities from equation (5) are predicted to diverge. Figure 4 details this dependence experimentally. If we vary the forcing such that the last smooth standing wave approaches the amplitude h_c —the value above which we observe bubble

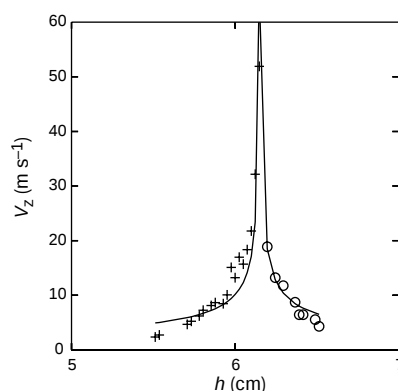


Figure 4 Dependence of jet velocity on surface wave height for $1.94 \text{ cm}^2 \text{ s}^{-1}$ fluid. The velocity of the jet tip depends strongly on the height of the last smooth standing wave before the singularity. The tip velocity increases rapidly as $h \rightarrow h_c$. The bubble formation (for $h > h_c$) requires energy, and the tip velocity decreases sharply with increasing bubble size. Near $h = h_c$, jet tip velocities exceeding 52 m s^{-1} have been observed. This velocity dependence is consistent with the Weber number We of the jet being derived from that of the last smooth wave. The solid line is the functional form $\rho v_z^2 (h - h_c) / \sigma = We$ where ρ is the fluid density, v_z is the jet tip velocity, and σ is the surface tension. Tip velocities are

found by observing the shadow of the jet tip. A vertical HeNe laser sheet passes horizontally through the tank through the central axis and shines on two phototransistors separated by a vertical distance of 2.7 cm . As the jet moves upward through the laser sheet, the curved droplet tip refracts the light and forms a shadow. Using a digital oscilloscope, the time delay between the appearance of the shadow on the two phototransistors is measured, and the velocity of the tip obtained. The high-viscosity fluid simplifies these measurements by damping out small-scale perturbations on the surface and creating smooth, vertical jets.

entrainment to first occur—the velocity peaks sharply, reaching values higher than 52 m s^{-1} . The entrainment of even very small bubbles decreases the jet velocity significantly.

We now analyse the surface profile of a subcritical collapsing depression—a near singularity. Each curve in Fig. 2a represents the shape of the surface profile at a given time before the singularity. We scale these data according to the similarity solution, using the location of the origin to which the profiles collapse vertically and the time of the singularity as parameters. Figure 2b shows the result of this scaling superimposed on the numerical solution. It is clear that the numerical solution is in good agreement with the observed surface profiles, indicating that the analysis captures the behaviour of the system. The collapse of the experimental data is consistent with a two-thirds scaling exponent, though for late times, an ultraviolet cutoff due to viscous effects causes increasing deviation from the self-similar form. When the fluid velocity becomes sufficiently large as the singularity is approached, viscous forces can no longer be neglected, and the above analysis would need to be extended to include rotation. Experimentally, overhang in the surface profile is seen to develop, suggesting that the fluid is not irrotational. As the fluid velocity is increased, the viscous ultraviolet cutoff also results in systematic rounding of the asymptotic cone. These viscous effects can be decreased by using a fluid with lower viscosity. However, in low-viscosity fluids, small-scale structures are observed on the standing waves before the collapse. Although we observe curvature collapse and jets in water (viscosity $0.01 \text{ cm}^2 \text{ s}^{-1}$), parasitic capillary waves and other perturbations complicate these surface profiles and make them difficult to analyse.

In most previous applications of similarity solutions, the original partial differential equations reduced to ordinary differential equations. However, in this work, our system reduces to partial differential equations—a more difficult, but tractable, situation^{9,10}. This method can now be applied to such problems as sheet-like jets erupting from fluid surfaces, as is seen in wave breaking on a cliff²⁰ in the gravity-dominated case. A power-law scaling for the form of the fluid jets produced in our experiments by the surface curvature collapse is also known²¹. One challenge remaining is to match the solution discussed here regarding the collapsing depression before the singularity with the known behaviour of the jet after the singularity. In addition, further study of the viscous ultraviolet

cutoff very close in time to the singularity is needed, and it would also be of interest to investigate the dependence of solutions on the Weber number. □

Received 25 March; accepted 11 November 1999.

- Shen, Y. R. *The Principles of Nonlinear Optics* (Wiley, New York, 1984).
- Weinberg, S. *Gravitation and Cosmology* (Wiley, New York, 1972).
- Newell, A. C. & Zakharov, V. E. Rough sea foam. *Phys. Rev. Lett.* **69**, 1149–1151 (1992).
- Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A neoproterozoic snowball earth. *Science* **281**, 1342–1346 (1998).
- Sarmiento, J. L. & Le Quere, C. Oceanic carbon dioxide uptake in a model of century-scale global warming. *Science* **274**, 1346–1350 (1996).
- Eggers, J. & Dupont, T. F. Drop formation in a one-dimensional approximation of the Navier-Stokes equation. *J. Fluid Mech.* **262**, 205–221 (1994).
- Eggers, J. Universal pinching of 3D axisymmetric free surface flow. *Phys. Rev. Lett.* **71**, 3458–3460 (1993).
- Brenner, M. et al. Breakdown of scaling in droplet fission at high Reynolds number. *Phys. Fluids* **9**, 1573–1590 (1997).
- Day, R. E., Hinch, E. J. & Lister, J. R. Self-similar capillary pinchoff of an inviscid fluid. *Phys. Rev. Lett.* **80**, 704–707 (1998).
- Keller, J. B. & Miksis, M. J. Surface tension driven flows. *SIAM J. Appl. Math.* **43**, 268–277 (1983).
- Hou, T. Y., Lowengrub, J. S. & Shelley, M. J. The long-time motion of vortex sheets with surface tension. *Phys. Fluids* **9**, 1933–1954 (1997).
- Chen, Y.-J. & Steen, P. H. Dynamics of inviscid capillary breakup: collapse and pinchoff of a film bridge. *J. Fluid Mech.* **341**, 245–267 (1997).
- Longuet-Higgins, M. S. Bubbles, breaking waves and hyperbolic jets at a free surface. *J. Fluid Mech.* **127**, 103–121 (1983).
- Faraday, M. On a peculiar class of acoustical figures, and on certain forms assumed by groups of particles on vibrating elastic surfaces. *Phil. Trans. R. Soc. Lond.* **121**, 299–340 (1831).
- Boulton-Stone, J. M. & Blake, J. R. Gas bubbles bursting at a free surface. *J. Fluid Mech.* **254**, 437–466 (1993).
- Longuet-Higgins, M. S. & Oguz, H. N. Critical microjets in collapsing cavities. *J. Fluid Mech.* **290**, 183–201 (1995).
- Longuet-Higgins, M. S. & Oguz, H. N. Critical jets in surface waves and collapsing cavities. *Phil. Trans. R. Soc. Lond. A* **355**, 625–639 (1997).
- Barenblatt, G. I. *Scaling, Self-similarity, and Intermediate Asymptotics* (Cambridge University Press, 1996).
- Zakharov, V. E. in *Breaking Waves*. (eds Banner, M. L. & Grisham, R. H. J.) (Springer, Berlin, 1992).
- Cooker, M. J. & Peregrine, D. H. Violent surface motion as near-breaking waves meet a wall. In *Breaking Waves*. (eds Banner, M. L. & Grisham, R. H. J.) (Springer, Berlin, 1992).
- Hogrefe, J. E. et al. Power-law singularities in gravity-capillary waves. *Physica D* **123**, 183–205 (1998).

Acknowledgements

We thank R. Rohde, M. Brenner, A. Bertozzi, J. Drake, T. Antonsen, E. Ott and J. Pyle for their input. D.P.L. is a Cottrell Scholar of the Research Corporation. This work was supported by the National Science Foundation.

Correspondence and requests for materials should be addressed to D.P.L. (e-mail: dpl@complex.umd.edu).

A single-photon detector in the far-infrared range

S. Komiyama*, O. Astafiev†, V. Antonov†‡, T. Kutsuwa* & H. Hirai*

* Department of Basic Science, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153, Japan

† CREST, Japan Science and Technology Corporation (JST), Kawaguchi, Saitama 332-0012, Japan

The far-infrared region (wavelengths in the range $10\text{ }\mu\text{m}$ – 1 mm) is one of the richest areas of spectroscopic research¹, encompassing the rotational spectra of molecules and vibrational spectra of solids, liquids and gases. But studies in this spectral region are hampered by the absence of sensitive detectors^{2–5}—despite recent efforts to improve superconducting bolometers⁶, attainable sensitivities are currently far below the level of single-photon detection. This is in marked contrast to the visible and near-infrared regions (wavelengths shorter than about $1.5\text{ }\mu\text{m}$), in which single-photon counting is possible using photomultiplier tubes. Here we report the detection of single far-infrared photons in the wavelength range 175 – $210\text{ }\mu\text{m}$ (6.0 – 7.1 meV), using a single-electron transistor consisting of a semiconductor quantum dot in high magnetic field. We detect, with a time resolution of a millisecond, an incident flux of 0.1 photons per second on an effective detector area of 0.1 mm^2 —a sensitivity that exceeds previously reported values by a factor of more than 10^4 . The sensitivity is a consequence of the unconventional detection mechanism, in which one absorbed photon leads to a current of 10^6 – 10^{12} electrons through the quantum dot. By contrast, mechanisms of conventional detectors^{2–6} or photon assisted tunnelling⁷ in single-electron transistors produce only a few electrons per incident photon.

We use a relatively large semiconductor quantum dot (QD), which is placed in a high magnetic field and is weakly coupled through tunnel barriers to electron reservoirs (Fig. 1a and b). The magnetic field is such that the lowest orbital Landau level, LL1, (with two opposite spin polarizations) is filled, while the first excited Landau level, LL2, is occupied with a small number of electrons. At the Fermi level, LL1 and LL2 form two compressible metallic regions, which correspond, respectively, to an “outer ring” and an “inner core” of the QD (Fig. 1a)⁸. The tunnelling probability between these two metallic regions is strongly suppressed by an incompressible insulating strip that separates them. When the electrochemical potential of the outer ring, μ_1 , lines up with that of the reservoirs, conductance resonance takes place because electrons can tunnel between the reservoirs and the outer ring. Although the inner core does not directly contribute to the transport, its charge strongly affects the transport by electrostatic coupling to the outer ring.

When a far-infrared (FIR) photon is absorbed by the QD upon cyclotron resonance (Fig. 1b), an electron (hole) created in the higher empty LL2 (the lower filled LL1) will rapidly give up its excess energy to the lattice and fall (climb up) to the inner core (outer ring). As the inner core is now negatively charged by $-e$ (the electron charge), the electrochemical potential of the outer ring has decreased by $-\Delta\mu_1 = -eC_2/(C_2C_1 + C_{12}C_1 + C_2C_{12})$, where C_i and C_{ij} are the capacitances formed between respective regions (Fig. 1a). This internal polarization causes the conductance-resonance peaks to shift by $-\Delta V_g \propto -\Delta\mu_1$ (Fig. 1d). This, together with the fact that the excited electron in the inner core has an extremely long lifetime owing to suppression of tunnelling, makes single FIR photon detection feasible.

In the experiments, a QD of mechanical size $700 \times 700\text{ nm}^2$ is fabricated on a GaAs/AlGaAs single heterostructure crystal (Fig. 1c). As extremely weak and well defined sources of FIR radiation, we use a GaAs two-dimensional electron gas (2DEG) Hall bar⁹ as well as an n-InSb device¹⁰. When the current I_{em} is passed through the devices in magnetic field B_{em} , they emit narrow-band cyclotron radiation centred at the cyclotron frequency $\omega_c = eB_{\text{em}}/m^*$ (refs 9, 10), where $m^* = 0.068 m_0$ and $m^* = 0.0014 m_0$ are, respectively, the effective masses of electrons in GaAs and InSb with the free electron mass m_0 . Two different arrangements have been applied for illumination. In one arrangement, a GaAs emitter is placed in the same mixing chamber as that of the QD sample. In the other arrangement, the emitters are placed in another superconducting solenoid ($T = 2\text{ K}$) located outside the mixing chamber (but still in the same cryostat), facilitating independent tuning of the radiation wavelength. In either arrangement, unwanted near- and mid-infrared radiation from the emitters, if any, is completely eliminated by filtering through silicon plates and black polyethylene films.

The black circles in Fig. 2 show the positions of the conductance peaks as a function of the control gate voltage V_g and magnetic field B taken without FIR illumination. As B is raised for a given trace, electrons are transferred sequentially from the inner core to the outer ring, while the total number of electrons is kept unchanged^{7,8}. Each time one electron leaves the inner core, the inner core is positively charged by $+e$, causing μ_1 to increase by $+\Delta\mu_1$. This process, exactly opposite to the one expected for the cyclotron-resonance excitation, yields a sequence of small jumps in the peak position by $\Delta V_g = +0.6\text{ mV}$ for each trace.

Single FIR photons have been detected in the B range of 3.4 – 4.15 T , where the inner LL2 contains from twenty electrons down to

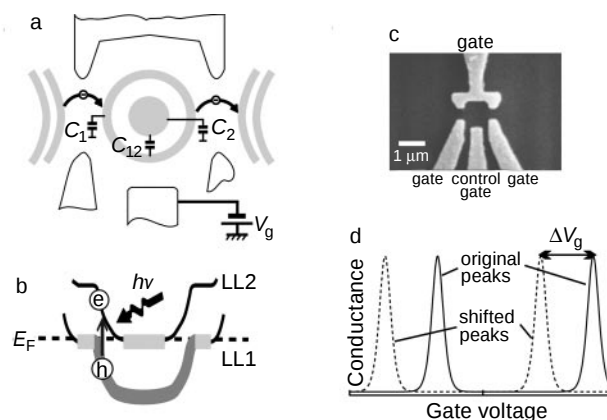


Figure 1 Single-electron transistor with a quantum dot (QD) in a magnetic field. The opposite spin states will not be distinguished in this description, as they are nearly degenerate under these experimental conditions. **a**, Diagram of the QD. The grey regions indicate metallic inner core and outer ring regions formed by the lowest two Landau levels (LLs). **b**, Energy spectra of the LLs in the QD. When an electron–hole pair is excited within the QD by absorption of a far-infrared (FIR) photon, the excited electron (hole) rapidly falls (climbs up) to the inner core (outer ring) of the QD to polarize the QD. **c**, Scanning electron micrograph of the QD ($700 \times 700\text{ nm}^2$), formed by laterally confining the 2DEG with negatively biased metal gates. The mobility and the sheet density of the 2DEG are, respectively, $\mu \approx 80\text{ m}^2\text{ V}^{-1}\text{ s}^{-1}$ and $n_s \approx 2.4 \times 10^{15}\text{ m}^{-2}$. The QD contains about 350 electrons. The metal gates and the leads, spanning a length of $100\text{ }\mu\text{m}$, serve as a dipole antenna that couples FIR radiation to the QD. The QD sample is placed within a mixing chamber of a dilution refrigerator. The magnetic field is applied normal to the plane of the QD with a superconducting solenoid. **d**, Diagram of Coulomb conductance peaks as a function of the control gate voltage V_g . In the single-electron transistor operation, the conductance resonance takes place when the electrochemical potential of the outer ring, μ_1 , lines up with that of the reservoirs. Upon excitation of the electron–hole pair in the QD, the conductance peak pattern (solid line) switches to a new pattern (dotted line) as a result of the induced polarization.

‡ Present address: Department of Physics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK.

one electron. Figure 3a shows a typical example obtained on one conductance peak at $T = 0.05$ K and $B = 3.67$ T. When very weak FIR illumination is turned on ($I_{em} = 2 \mu\text{A}$), the original conductance resonance is occasionally switched off, while the finite conductance shows up as a few spikes distributed over a V_g range shifted towards the more negative direction of V_g . When the intensity of illumination increases ($I_{em} = 3.5 \mu\text{A}$), the original conductance resonance nearly vanishes, while conductance in the shifted V_g range grows to form a dense array of spikes defining a distinct envelope of another resonance line centred at the shifted position by $\Delta V_g = -0.6$ mV. Detailed studies reveal that the peak position of the new conductance resonance ($\Delta V_g = -0.6$ mV) is independent of the FIR intensity and that the disappearance of the original conductance resonance (integrated over time) is exactly compensated by the appearance of the new conductance resonance (integrated over time) at any level of FIR intensity. These findings indicate that this new conductance resonance is due to the state of the QD in which one extra electron is excited by the cyclotron resonance in the inner core. Finite conductance takes the form of spikes, also in a V_g range shifted by $\Delta V_g = -1.2$ mV at $I_{em} = 3.5 \mu\text{A}$, which corresponds to a doubly excited state when two electron-hole pairs are excited within the QD.

The shifted conductance resonance, recognized as the envelope of conductance spikes under the FIR illumination, shows neither width broadening nor amplitude reduction compared to the original resonance line. This indicates that the effective electron temperature is not appreciably elevated (not by more than 10 mK) by the FIR illumination. We have studied the influence of elevated lattice temperature up to 0.4 K in additional experiments, and

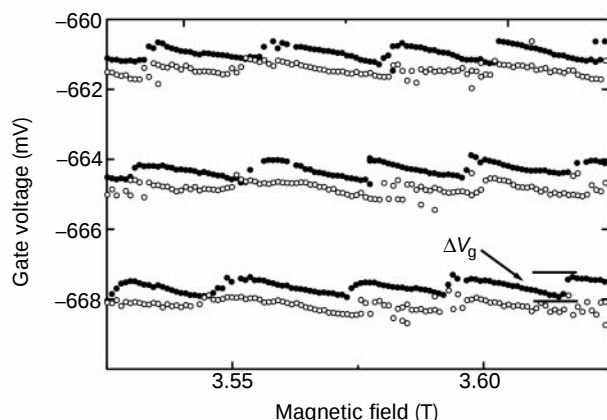


Figure 2 Positions of the Coulomb conductance peaks in the plane of V_g versus B . Data shown for observations without FIR illumination (black circles) and with FIR illumination (white circles; $I_{em} = 3.5 \mu\text{A}$). Without illumination, each trace maps out the boundary conditions at which the total number of electrons on the QD increases by one as V_g is increased. For a given trace, electrons are transferred sequentially from the inner core (the first excited Landau level, LL2) to the outer ring (the lowest orbital Landau level, LL1) as B is raised, while the total number of electrons is kept unchanged⁸. Each time one electron leaves the inner core, the inner core is charged by $+e$ and causes a stepwise discontinuous shift of the peak position by $\Delta V_g \approx 0.6$ mV. This intra-QD charge transfer repeats approximately each time one more flux quantum is added to LL1. The average period of the steps, $\Delta B = 20$ mT, suggests an effective dot area of 500 nm^2 , a size which is consistent with the dimensions for the QD. Under illumination, the conductance resonance occurs at shifted V_g positions (white circles), which form the lines that are separated from the non-excited traces in the negative direction by an amplitude approximately equal to the step height ($\Delta V_g \approx -0.6$ mV) of the non-excited traces. For certain ranges of B , conductance resonance is seen to occur also at the doubly shifted positions ($\Delta V_g \approx -1.2$ mV). Such B ranges correspond to the regime where the lifetime of excited states are especially long and the doubly excited states are visible. All the data points in this figure are taken with a time constant much larger than the average time interval of the conductance switching described in Fig. 3. The arrangement for FIR illumination is the same as that for Fig. 3a.

confirmed that no conductance switches as observed here are induced. Thermal activation is thus irrelevant to the phenomena.

Similar effects of FIR illumination are observed on each conductance peak in the B range of 3.4–4.15 T. White circles in Fig. 2 trace the V_g positions of the new conductance-resonance peaks under illumination ($I_{em} = 3.5 \mu\text{A}$). The traces primarily form the lines that are shifted from the non-excited traces (black circles) in the negative direction of V_g by the same amplitude ($\Delta V_g \approx -0.6$ mV) as that of the steps exhibited by the non-excited

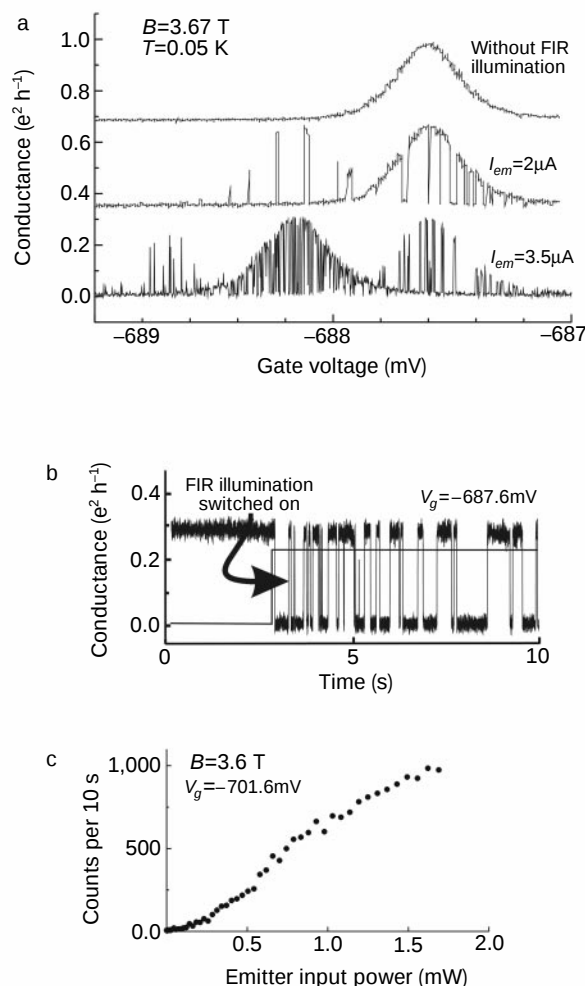


Figure 3 Conductance resonance and state switching. **a**, Conductance-resonance curve with and without far-infrared illumination. A GaAs-emitter is placed within the mixing chamber, where B_{em} is larger than B for the QD, by about 3%, in order to meet the plasma-shifted cyclotron resonance of the QD (ref. 11). The incident FIR power on the effective antenna area, $100 \mu\text{m}^2$, at the position of the QD sample, is less than ~ 0.003 fW ($\sim 10^4$ photons per second, in terms of photon flux) at $I_{em} = 3.5 \mu\text{A}$. The time constant of measurements is 1 ms and the gate voltage is scanned over 3 min for each curve. Without FIR illumination, the conductance resonance yields a smooth curve (shown at the top). With FIR illumination ($I_{em} = 2.5$ and $3.5 \mu\text{A}$), the smooth curve is replaced by a sequence of frequently switching spikes. The envelope of the spikes reveals a new conductance resonance at the position shifted by $\Delta V_g = -0.6$ mV from the original peak position. Signals due to the doubly excited state with two extra electrons in the inner core can be also recognized at the position shifted by $\Delta V_g = -1.2$ mV. **b**, Time-trace representation of the data of **a** with V_g fixed at the original peak position and $I_{em} = 3.5 \mu\text{A}$. The arrangement of FIR illumination is the same as that for **a**. **c**, The number of switching-off events counted over ten seconds is plotted as a function of the electrical input fed to the GaAs emitter. The emitter is outside the mixing chamber, where B_{em} is set to be at the resonance position. The input power of 0.1 mW roughly corresponds to the FIR intensity of $I_{em} = 3.5 \mu\text{A}$ for **a** and **b**. (See the legend of Fig. 4.) The control gate voltage V_g for the QD is fixed at the original conductance-peak position.

traces. This supports our interpretation that the new conductance resonance arises from the state containing one extra electron in the inner core. Also, several data points are located at the doubly shifted positions ($\Delta V_g \approx -1.2$ mV), corresponding to the doubly excited state.

When the gate voltage is fixed at the original peak position of a given conductance resonance, the effect is of telegraph-like switches between two conductance states, as shown in Fig. 3b. Each switching-off event corresponds to an individual process of photon absorption, and each switching-on event to recombination of an excited electron-hole pair. As shown in Fig. 3c, the rate of switching-off events increases with increasing the FIR intensity at a low-excitation level, while the average time duration over which the switched-off state is maintained (the lifetime of the excited state) is kept unchanged. The rate of switching-off events levels off and starts to decrease at high levels of excitation where the average time interval between successive photon absorptions becomes shorter than the lifetime of the excited state (data not shown). If an additional electron-hole pair is excited already before a previously created electron-hole pair recombines, the QD undergoes transition to the doubly excited state, with no observable switch.

The conductance switches described here take place only when the FIR radiation is in resonance to the inter-Landau-level excitation energy of the QD, as demonstrated in Fig. 4, where the number of switching-off events of conductance counted over ten seconds is recorded as a function of the FIR photon energy (B_{em}). The absence of background signal shows that conductance switches without illumination are negligible. The magnetic fields, B_{em} , at which the respective emitters yield the largest number of counts correspond to the "magneto-plasma resonance" frequency of the QD (ref. 11) that is slightly higher than the cyclotron-resonance frequency of the bulk 2DEG (by about 3%). The spectral line width is determined by that of the radiation from the emitters. Hence, the true spectral response

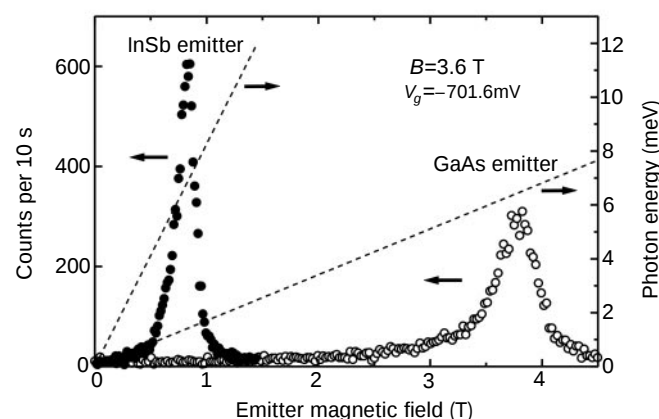


Figure 4 Excitation spectra of the quantum dot. The number of switching-off events of conductance over ten seconds, occurring when V_g is fixed at the original peak position, is shown versus B_{em} . The measurements are repeated while B_{em} is increased in steps of 20 mT. The FIR radiation is generated by a GaAs 2DEG emitter ($\mu_H = 20$ mV s $^{-1}$ and $N_s = 1.8 \times 10^{15}$ m $^{-2}$) and an n-InSb emitter placed outside the mixing chamber. The radiation is guided through a heat-anchored metal light pipe, 80 cm in length and with a 3-mm bore, using sufficient filtering, and introduced into the mixing chamber through a sealed silicon window. The corresponding photon energies are indicated by the broken lines. To keep the intensity of FIR illumination unchanged at a sufficiently low level over the entire spectral range, the electrical input to the emitters is kept to be constant at 0.5 mW: the FIR power effectively incident on the QD is estimated to be of the order of 0.3 fW. The magnetic field ($B = 3.6$ T) for the QD is the same as that for Fig. 3c but different from that for Fig. 3a and b. It follows that the lifetime of the excited state of the QD is about 5 ms, the same as that for Fig. 3c but substantially shorter than that for Fig. 3a and b. This condition of a relatively short lifetime is intentionally chosen for the practical purpose of measurements in order to avoid a too-long integration time for accumulating a sufficient photon count.

of the QD is probably sharper than the resonance lines presented here. These results indicate that the conductance switches that we observed here arise from individual events of single FIR photon absorption by the QD.

We briefly discuss the lifetime of the excited state of the QD. As a general trend, the lifetime markedly increases from the order of 1 ms up to the order of 1,000 s as B increases from 3.4 T to 4.0 T. Above 4.0 T, it sharply decreases. This overall behaviour is superposed with sharp saw-tooth-like oscillations that synchronize with the step structure of the ground-state peak position shown in Fig. 2. We interpret these features by considering the size evolution of the inner core and the outer ring of the QD with increasing B (ref. 12).

Thus, we have achieved single-photon detection in the wavelength range 175–210 μ m (6.0–7.1 meV) restricted by the magnetic-field range (3.4–4.15 T) where the lifetime of the excited state exceeds the time constant of measurements (1 ms). This can be expanded by faster measurements—the intrinsic limit on the speed of the single-electron transistor is greater than 10 GHz (ref. 13). In practice, using a high-frequency single-electron transistor¹³ will improve the time resolution up to a 10-MHz range. We have also successfully observed conductance switches owing to the single FIR photon absorption up to $T = 0.4$ K; the upper bound is limited by the charging energy of the QD (ref. 14) (about 0.4 meV). Fabricating QDs on a narrow 2DEG mesa structure with minimum use of nearby metal gates may expand the upper bound to $T \approx 0.8$ K; this could make possible the use of more convenient 3 He refrigerators. Very few conductance switches are seen if the device is not illuminated: in an optimum B range, the dark signal, in terms of the switch rate, was as small as 0.001 s $^{-1}$ at $T = 0.05$ K. Finally, the quantum yield, or the ratio of the detected photon number to the incident photon flux on the effective antenna area is roughly estimated to be 1%. The quantum yield crucially depends on the architecture of the optics, which has not been optimized and may be significantly improved. Here, the effective noise equivalent power (NEP) reaches the order of 10^{-22} W Hz $^{-1/2}$ in the optimum B range, which exceeds other FIR detectors reported in the literature by a factor of more than 10^4 . □

Received 27 July; accepted 6 December 1999.

- van Dishoeck, E. F. & Helmich, F. P. Scientific drivers for future high-resolution far-infrared spectroscopy in space. *Proc. Eur. Space Agency Symp.* **388**, 3–12 (1996).
- Fujiwara, M. & Hiromoto, N. Ge:Ga far infrared photoconductor with a low Ga concentration of 1×10^{14} cm $^{-3}$. *Jpn J. Appl. Phys.* **36**, 4262–4266 (1997).
- Kawamura, J. *et al.* Low-noise NbN lattice-cooled superconducting hot-electron bolometric mixers at submillimeter wavelengths. *Appl. Phys. Lett.* **70**, 1619–1621 (1997).
- Tucker, J. R. & Feldman, M. J. Quantum detection at millimeter wavelengths. *Rev. Mod. Phys.* **57**, 1055–1113 (1985).
- Richards, P. L. Bolometers for infrared and millimeter waves. *J. Appl. Phys.* **76**, 1–24 (1994).
- Schoelkopf, R. J., Moseley, S. H., Stahl, C. M., Wahlgreen, P. & Delsing, P. A concept for a submillimeter-wave single-photon counter. *IEEE Trans. Appl. Supercond.* **9**, 2935–2939 (1999).
- Kouwenhoven, L. P. *et al.* in *Mesoscopic Electron Transport* (eds Sohn, L. L. *et al.*) 105–214 (Kluwer Academic, Dordrecht, The Netherlands, 1997).
- van der Vaart, N. C. *et al.* Time-resolved tunneling in the quantum Hall regime. *Phys. Rev. B* **55**, 9746–9756 (1997).
- Kawano, Y., Hisanaga, Y. & Komiyama, S. Cyclotron emission from quantized Hall devices: injection of nonequilibrium electrons from contacts. *Phys. Rev. B* **59**, 12537–12546 (1999).
- Gornik, E. *et al.* Fourier transform spectroscopy of Landau radiation in n-InSb. *Infrared Phys.* **16**, 285–291 (1976).
- Bollweg, K. *et al.* Detection of compressible and incompressible states in quantum dots and antidots by FIR-spectroscopy. *Phys. Rev. Lett.* **76**, 2774–2777 (1996).
- Stopa, M., Bird, J. P., Ishibashi, K., Aoyagi, Y. & Sugano, T. Giant backscattering magnetoresistance resonance and Quantum Dot electronic structure. *Phys. Rev. Lett.* **76**, 2145–2148 (1996).
- Schoelkopf, R. J., Wahlgreen, P., Kozhevnikov, A. A., Delsing, P. & Prober, D. E. The radio-frequency single-electron transistor (RF-SET): a fast and ultrasensitive electrometer. *Science* **280**, 1238–1241 (1998).
- Averin, D. V. & Likharev, K. K. Coulomb blockade of single-electron tunneling, and coherent oscillations in small tunnel junctions. *J. Low Temp. Phys.* **62**, 345–373 (1986).

Acknowledgements

We thank D. Austing and M. Stopa for helpful discussion. This work was supported by the Core Research for Evolutional Science and Technology (CREST) of the Japan Science and Technology Corporation (JST).

Correspondence and requests for materials should be addressed to S.K. (e-mail: csusumu@ASone.c.u.-tokyo.ac.jp).

Efficient organic photovoltaic diodes based on doped pentacene

J. H. Schön^{*†}, Ch. Kloc^{*}, E. Bucher^{*†} & B. Batlogg^{*}

^{*} Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974-0636, USA

[†] University of Konstanz, Faculty for Physics, P.O. Box X916, D-78457 Konstanz, Germany

Recent work on solar cells based on interpenetrating polymer networks^{1–3} and solid-state dye-sensitized devices⁴ shows that efficient solar-energy conversion is possible using organic materials. Further, it has been demonstrated that the performance of photovoltaic devices based on small molecules can be effectively enhanced by doping the organic material with electron-accepting molecules⁵. But as inorganic solar cells show much higher efficiencies, well above 15 per cent, the practical utility of organic-based cells will require their fabrication by lower-cost techniques, ideally on flexible substrates. Here we demonstrate efficiency enhancement by molecular doping in Schottky-type photovoltaic diodes based on pentacene—an organic semiconductor that has received much attention as a promising material for organic thin-film transistors^{6–8}, but relatively little attention for use in photovoltaic devices^{9,10}. The incorporation of the dopant improves the internal quantum efficiency by more than five orders of magnitude and yields an external energy conversion efficiency as high as 2.4 per cent for a standard solar spectrum. Thin-film devices based on doped pentacene therefore appear promising for the production of efficient ‘plastic’ solar cells.

Schottky-type photovoltaic cells (see Fig. 1a) were fabricated based on vapour-grown pentacene single crystals¹¹. The crystals (in

about 1 μm thickness) were doped in a solution of acetonitrile and iodine or bromine. The dopant is incorporated in the crystal between the molecular pentacene layers¹². An ohmic, transparent indium–tin–oxide (ITO) layer and evaporation of an additional silver contact finger. The aluminium back contact, which forms the Schottky barrier, was prepared by thermal evaporation. The photovoltaic performance of such a device under a standard solar irradiation of 100 mW cm^{-2} (air mass AM1.5) is shown in Fig. 1b. For bromine- and iodine-doped devices we measure an open-circuit voltage V_{oc} of 970 and 955 mV, a short-circuit current density of 5.3 and 4.5 mA cm^{-2} , and a fill factor, which is the ratio of the maximum power of the photovoltaic device to the product of open-circuit voltage and short-circuit current, of 0.47 and 0.46. This results in an overall energy conversion efficiency η of 2.4% and 1.9% at AM1.5 conditions. (Electron-beam-induced current measurements clearly revealed that only the contact area was active. Hence, it appears reasonable to assume that no carriers are collected from the uncontacted part of the surface; the short-circuit current density is therefore calculated using the contacted area of the device.) A maximum efficiency of 2.6% and 2.1% was found for 20 mW cm^{-2} irradiation intensity, which is ascribed to lower resistive losses at lower illumination intensities.

The incorporation of bromine and iodine as dopant is beneficial for the photovoltaic properties in several ways: (1) the conductivity of pentacene is increased, (2) the optical absorption edge is shifted to 1.4 eV, and (3) the quantum efficiency is enhanced. Altogether, the device performance increases by more than five orders of magnitude (undoped device η below $10^{-5}\%$, see Fig. 2b).

The conductivity of pentacene can be modified in a controlled way through doping between $10^{-12} \text{ S cm}^{-1}$ for as-grown crystals¹³ and 100 S cm^{-1} (ref. 14) for highly doped materials. The in-plane charge-carrier (hole) mobility changes only slightly over this large range of doping from 1.9 to $0.3 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ (Fig. 2a); thus, the loss

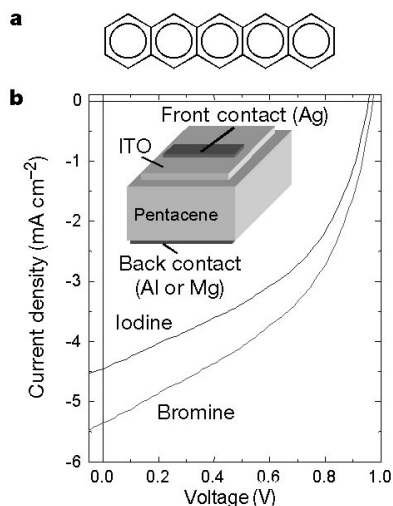


Figure 1 Chemical structure and current–voltage characteristics. **a**, Chemical structure of a pentacene molecule. **b**, Current–voltage characteristics of an iodine-doped and a bromine-doped pentacene solar cell under simulated solar illumination (AM1.5, 100 mW cm^{-2}). The energy-conversion efficiency of the iodine-doped (bromine-doped) device is 1.9% (2.4%) with an open-circuit voltage of 955 mV (970 mV), a short-circuit current density of 4.6 mA cm^{-2} (5.3 mA cm^{-2}), and a fill factor of 46% (47%). The contacted (illuminated) areas of the devices were 5.6 (7.2) and 6.7 (8.5) mm^2 for the iodine- and bromine-doped cell, respectively. The schematic device structure is shown in the inset. The pentacene is grown from the vapour phase and doped in an acetonitrile/iodine solution. The back contact is prepared by thermal evaporation and the transparent indium–tin–oxide (ITO) front electrode is sputtered. In addition, a silver contact finger is evaporated onto the ITO layer.

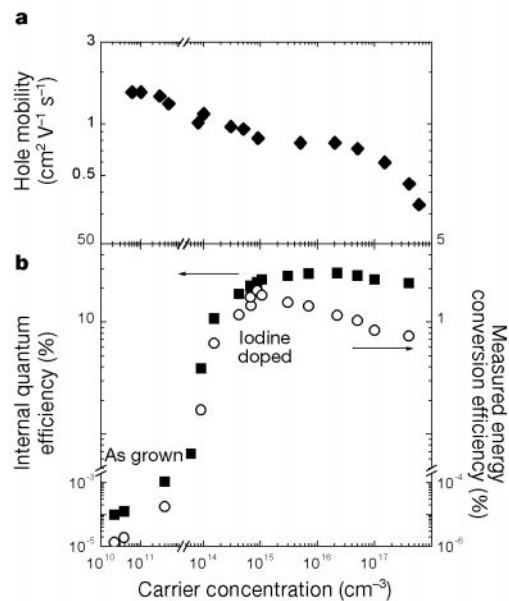


Figure 2 Charge-carrier mobility, internal quantum efficiency (IQE) and photovoltaic efficiency as functions of carrier concentration. **a**, The charge carrier (hole) mobility μ (black diamonds) in as-grown and iodine-doped pentacene as a function of carrier concentration p . Although p is varied by eight orders of magnitude, μ changes only slightly from 1.9 to $0.3 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. **b**, IQE and η as functions of p . IQE (black squares) and η (white circles) increase by more than five orders of magnitude due to iodine doping. A maximum IQE of 27.5% and a maximum η of 1.9% are reached.

due to the series resistance of the device can be dramatically reduced. However, the anisotropy, that is, the ratio of the mobility for charge transport perpendicular to the molecular planes to the mobility for in-plane transport, increases very rapidly for carrier concentrations exceeding 10^{14} cm^{-3} and reaches values of up to 10^8 at the highest dopant densities, where iodine or bromine intercalates between the pentacene layers. Furthermore, the absorption edge shifts to lower energies with the incorporation of iodine or bromine, and produces a spectral response that is well matched to the solar spectrum. The absorption edge of 1.4 eV is similar to that of GaAs and is optimal for single junction devices¹⁵. In addition, the internal quantum efficiency (IQE) follows the absorption coefficient (Fig. 3) at small energies and shows a maximum of 36% and 27.5% at 1.9 eV for bromine- and iodine-doped photovoltaic diodes, respectively. The low-energetic part can be modelled using the absorption coefficient α , the depletion layer width W (which is determined by capacitance–voltage C – V measurements) and the minority carrier diffusion length L_n . The quantum efficiency is given by:

$$\text{IQE} = 1 - \frac{\exp(-\alpha W)}{1 + \alpha L_n} \quad (1)$$

Using the measured IQE and equation (1) a diffusion length L_n as large as 125 nm is obtained in iodine-doped pentacene. It is well known that related polyacenes, like naphthalene and anthracene, exhibit high electron mobilities ($1\text{--}2 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$; ref. 16), which is an essential prerequisite for a long minority carrier diffusion length. Similar electron mobilities can be expected also in pentacene. Moreover, doping increases the internal quantum efficiency (number of charge carriers per photon) from $10^{-4}\%$ in as-grown pentacene to 27.5% in iodine-doped devices. This enhancement can be explained by a similar process to that seen in doped merocyanine photovoltaic devices^{5,17}. The incoming photon is absorbed in the active layer and an exciton is formed (Fig. 4a). The exciton diffuses to the vicinity of a dopant and an excited pentacene–iodine–pentacene complex is formed (Fig. 4b), where the negative charge is partly transferred to the dopant molecule (Fig. 4c). Subsequently, the hole and the electron separate and move apart by the built-in electric field resulting in an enhanced photocurrent (Fig. 4d).

As function of carrier concentration p the IQE increases by many orders of magnitude, peaks near $p \approx 10^{16} \text{ cm}^{-3}$, and decreases

slightly for higher p . Figure 3 shows the IQE and η (AM1.5, 100 mW cm^{-2}) as function of the carrier concentration p , which was determined by capacitance–voltage measurements. The small decrease at high p may be attributed to a stronger scattering and recombination of photo-generated carriers at higher dopant concentrations caused by the increased number of structural imperfections. The dependence of η is similar and reaches a maximum for $p \approx 10^{15} \text{ cm}^{-3}$. Voltage-dependent measurements of the quantum efficiency revealed that the slight decrease of η for $p > 10^{15} \text{ cm}^{-3}$ is related to an incomplete carrier collection, as the sample thickness exceeds the depletion layer width. (Space-charge-limited current measurements revealed trap densities of the order of 10^{15} cm^{-3} in the case of doped pentacene, which is comparable to the optimum dopant concentration. As a result, we assume that neither the exciton nor the charge-carrier diffusion length will exceed a few hundred nanometres. This is in line with the value estimated by IQE measurements.)

We propose that the dramatic increase of the IQE for p exceeding 10^{14} cm^{-3} is due to the fact that at this dopant concentration the average distance between the dopants becomes comparable to the exciton diffusion length. Thus, the charge separation becomes more efficient.

Here we have optimized the efficiency by optimizing only the doping level. Additional improvements may be anticipated when additional device parameters are varied. A first obvious step involves optimizing the thickness of the pentacene layer. As the depletion-layer width W decreases with increasing p ($W \propto \sqrt{1/p}$), not all generated carriers may be separated by the electric field resulting in a sub-optimal short-circuit current. Thickness adjustments can best be implemented by the use of thin-film devices. Pentacene thin films have shown high quality and high charge carrier mobilities^{6–8,14}. Apparently the electronic levels caused by the grain boundaries are shallow enough not to impair the mobilities at room temperature⁷. In this way the thickness of the device and the depletion-layer width can be optimized for any given dopant concentration. Furthermore, reduced thickness and increased carrier density diminishes the series resistance loss leading to a higher short-circuit density and fill factor. A further reduction of the series resistance might be expected if the pentacene molecules could be aligned in a thin film parallel to the substrate; this has been shown recently for octithiophene-based photovoltaic devices¹⁸. In this case,

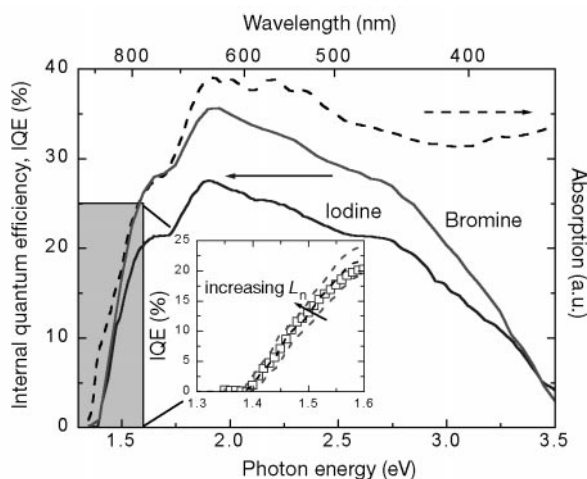


Figure 3 IQE and optical absorption of doped pentacene photovoltaic cells as function of energy. The IQE follows the absorption at low energies and reaches a maximum value of 27.5% and 36% for iodine- and bromine-doped devices, respectively. Inset, magnification of the IQE at the absorption edge and a fit according to equation (1) using different values of the minority carrier diffusion length L_n (100, 125, and 150 nm). L_n is estimated to be 125 nm.

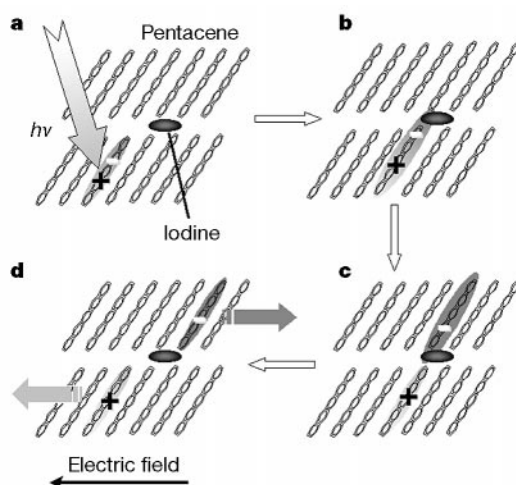


Figure 4 Charge-carrier generation and separation in iodine-doped pentacene. **a**, The incoming photon ($h\nu$) is absorbed and an exciton is formed. **b**, **c**, This exciton diffuses to a dopant site (**b**) forming a complex with the dopant and the neighbouring pentacene molecules (**c**). **d**, Charge is transferred to the dopant and, finally, the charges are separated in the built-in electric field of the device and transported to the contacts.

the in-plane mobility of the charge carriers would determine the device characteristics rather than the mobility perpendicular to the molecular planes, as in the case of single crystal cells. At optimum doping ($p \approx 10^{15} \text{ cm}^{-3}$) the anisotropy is of the order of 10^2 or 10^3 .

To enhance the open-circuit voltage we have used other low-work-function metals (like magnesium and calcium) as back contacts and found an enhancement of V_{oc} by up to 250 mV resulting in a maximum V_{oc} of 1,200 mV. At this point, the quality of these contacts needs further work to reduce losses caused by a low shunt resistance. Initial studies of device stability indicate no degradation during the first hours of operation. However, after 100 hours a loss in efficiency of approximately 10% is observed, which we may ascribe to interaction of iodine or bromine and aluminium at the back contact. In a first step to improve the long-term stability of these photovoltaic devices, dopants will have to be found that do not interact with the back-contact material. Then the intrinsic stability of the doped pentacene against photo-induced degradation will need to be studied.

The present results demonstrate that respectable photovoltaic performance can be achieved in a simply doped Schottky-diode structure based on organic semiconductors. Furthermore, the use of thin films, where thickness, conductivity, and doping ratio can be optimized, or the combination of doped pentacene with a suitable n-type material in a two-layer device^{19–21} show potentially even higher efficiencies. These first research results and progress in polymeric devices might open new perspectives for efficient 'plastic' solar cells. □

Methods

Crystal growth and doping

Pentacene single crystals are grown by horizontal physical vapour transport in a stream of hydrogen. The volatilization zone is heated to approximately 300 °C and the growth zone is kept at 200–280 °C. Trap and dopant densities as low as 10^{13} and 10^{11} cm^{-3} were measured using space-charge-limited current measurements. Iodine or bromine doping was performed by immersing the crystals into a solution of iodine or bromine in acetonitrile. The dopant density is adjusted by the iodine or bromine concentration of the solution.

Device fabrication

An approximately 300 nm thick, transparent ITO window layer was deposited onto a freshly cleaved (about 1 μm thick), doped single crystal by radio frequency magnetron sputtering (Leybold Z400 sputtering system). The typical conductivity of this layer was about 2000 S cm^{-1} . In addition a silver-contact finger was deposited onto the window layer in order to reduce series resistance losses. The Schottky contact (Al/Mg, 100 nm thick) was thermally evaporated. Typical cell areas are in the range 2–5 mm^2 .

Current–voltage characteristics

Current–voltage curves are measured in air at room temperature using a Keithley 238 Source-Measure-Unit. A commercial solar simulator is used as illumination source (Steuernagel, solar constant 575). Its intensity is adjusted using a calibrated and certified GaAs solar cell (ISE Institute für Solare Energieforschung, Freiburg). The current density is not numerically adjusted for the spectral mismatch between GaAs and doped pentacene.

Received 1 June; accepted 25 November 1999.

- Halls, J. J. M. *et al.* Efficient photodiodes from interpenetrating polymer networks. *Nature* **376**, 498–500 (1995).
- Yu, G., Gao, J., Hummelen, J. C., Wudl, F. & Heeger, A. J. Polymer photovoltaic cells: Enhanced efficiencies via a network of internal donor-acceptor heterojunctions. *Science* **270**, 1789–1791 (1995).
- Granström, M. *et al.* Laminated fabrication of polymeric photovoltaic diodes. *Nature* **395**, 257–260 (1998).
- Bach, U. *et al.* Solid-state dye-sensitized mesoporous TiO_2 solar cells with high photon-to-electron conversion efficiencies. *Nature* **395**, 583–585 (1998).
- Chamberlain, G. A., Cooney, P. J. & Dennison, S. Photovoltaic properties of merocyanine solid-state photocells. *Nature* **289**, 45–47 (1981).
- Lin, Y.-Y., Gundlach, D. J., Nelson, S. F. & Jackson, T. N. Pentacene-based organic thin-film transistors. *IEEE Trans. Electr. Dev.* **44**, 1325–1331 (1997).
- Nelson, S. F., Lin, Y.-Y., Gundlach, D. J. & Jackson, T. N. Temperature-independent transport in high-mobility pentacene transistors. *Appl. Phys. Lett.* **72**, 1854–1856 (1998).
- Dimitrakopoulos, C. D., Purushothaman, S., Kymissis, J., Callegari, A. & Shaw, J. M. Low voltage organic transistors on plastic comprising high dielectric constant gate insulators. *Science* **283**, 822–824 (1999).
- Signerski, R., Jarosz, G. & Godlewski, J. Photoelectric properties of heterojunctions formed from di-(pyridyl)-perylene-tetracarboxylic diimide and copper phthalocyanine or pentacene. *Synth. Met.* **94**, 135–137 (1998).

- Videlot, C., Fichou, D. & Garnier, F. Photovoltaic cells based on organic semiconductors. *J. Chim. Phys.* **95**, 1335–1338 (1998).
- Laudise, R. A., Kloc, Ch., Simpkins, P. G. & Siegrist, T. Physical vapor growth of organic semiconductors. *J. Cryst. Growth* **187**, 449–454 (1998).
- Minikata, T., Imai, H., Ozaki, M. & Saco, K. Structural studies on highly ordered and highly conductive thin films of pentacene. *J. Appl. Phys.* **72**, 5220–5225 (1992).
- Schön, J. H., Kloc, Ch., Laudise, R. A. & Batlogg, B. Electrical properties of single crystals of rigid rodlike conjugate molecules. *Phys. Rev. B* **58**, 12952–12957 (1998).
- Minakata, T., Nagoya, I. & Ozaki, M. Highly ordered and conducting thin films of pentacene doped with iodine vapor. *J. Appl. Phys.* **69**, 7354–7356 (1991).
- Hovel, H. J. *Semiconductors and Semimetals* Vol. 11, *Solar Cells* 71–92 (Academic, New York, 1975).
- Karl, N. in *Semiconductors* (eds Madelung, O., Schulz, M. & Weiss, H.) Subvolume 17i, 106–218 (Landolt-Börnstein, New Series Vol. 17, Springer-Verlag, Berlin, 1985).
- Chamberlain, G. A. Depletion layer studies and carrier photogeneration in doped merocyanine photovoltaic cells. *J. Appl. Phys.* **53**, 6262–6269 (1982).
- Videlot, C. & Fichou, D. Influence of molecular orientation on the photovoltaic properties of octithiophene. *Synth. Met.* **102**, 885–888 (1999).
- Tang, C. W. Two-layer organic photovoltaic cell. *Appl. Phys. Lett.* **48**, 183–185 (1986).
- Wöhrl, D. *et al.* Investigations of n/p junction photovoltaic cells of perylene-tetracarboxylic acid diimides and phthalocyanines. *J. Mater. Chem.* **5**, 1819–1829 (1995).
- Wöhrl, D. & Meissner, D. Organic solar cells. *Adv. Mater.* **3**, 129–138 (1991).

Acknowledgements

We thank E. Chandross, Z. Bao, H. Katz and A. Dodabalapur for helpful discussions. J.H.S. gratefully acknowledges financial support by the 'Deutsche Forschungsgemeinschaft'.

Correspondence and requests for materials should be addressed to J.H.S. (e-mail: hendrik@lucent.com).

Rainfall and drought in equatorial east Africa during the past 1,100 years

Dirk Verschuren^{*†}, Kathleen R. Laird[‡] & Brian F. Cumming[‡]

^{*} Limnological Research Center, University of Minnesota, Minneapolis, Minnesota 55455, USA

[†] Department of Biology, University of Gent, Ledeganckstraat 35, B-9000 Gent, Belgium

[‡] Paleocological Environmental Assessment and Research Laboratory, Department of Biology, Queen's University, Kingston, Ontario K7L 3N6, Canada

Knowledge of natural long-term rainfall variability is essential for water-resource and land-use management in sub-humid regions of the world. In tropical Africa, data relevant to determining this variability are scarce because of the lack of long instrumental climate records and the limited potential of standard high-resolution proxy records such as tree rings and ice cores^{1–3}. Here we present a decade-scale reconstruction of rainfall and drought in equatorial east Africa over the past 1,100 years, based on lake-level and salinity fluctuations of Lake Naivasha (Kenya) inferred from three different palaeolimnological proxies: sediment stratigraphy and the species compositions of fossil diatom and midge assemblages. Our data indicate that, over the past millennium, equatorial east Africa has alternated between contrasting climate conditions, with significantly drier climate than today during the 'Medieval Warm Period' (~AD 1000–1270) and a relatively wet climate during the 'Little Ice Age' (~AD 1270–1850) which was interrupted by three prolonged dry episodes. We also find strong chronological links between the reconstructed history of natural long-term rainfall variation and the pre-colonial cultural history of east Africa⁴, highlighting the importance of a detailed knowledge of natural long-term rainfall fluctuations for sustainable socio-economic development.

The partially submerged Crescent Island crater (CIC) basin in Lake Naivasha, Kenya, is unique for equatorial east Africa in combining the climatic sensitivity of a shallow Rift Valley lake with the wind-sheltered sedimentation regime of a deep crater lake⁵. Naivasha is a freshwater lake with subsurface outflow and a short

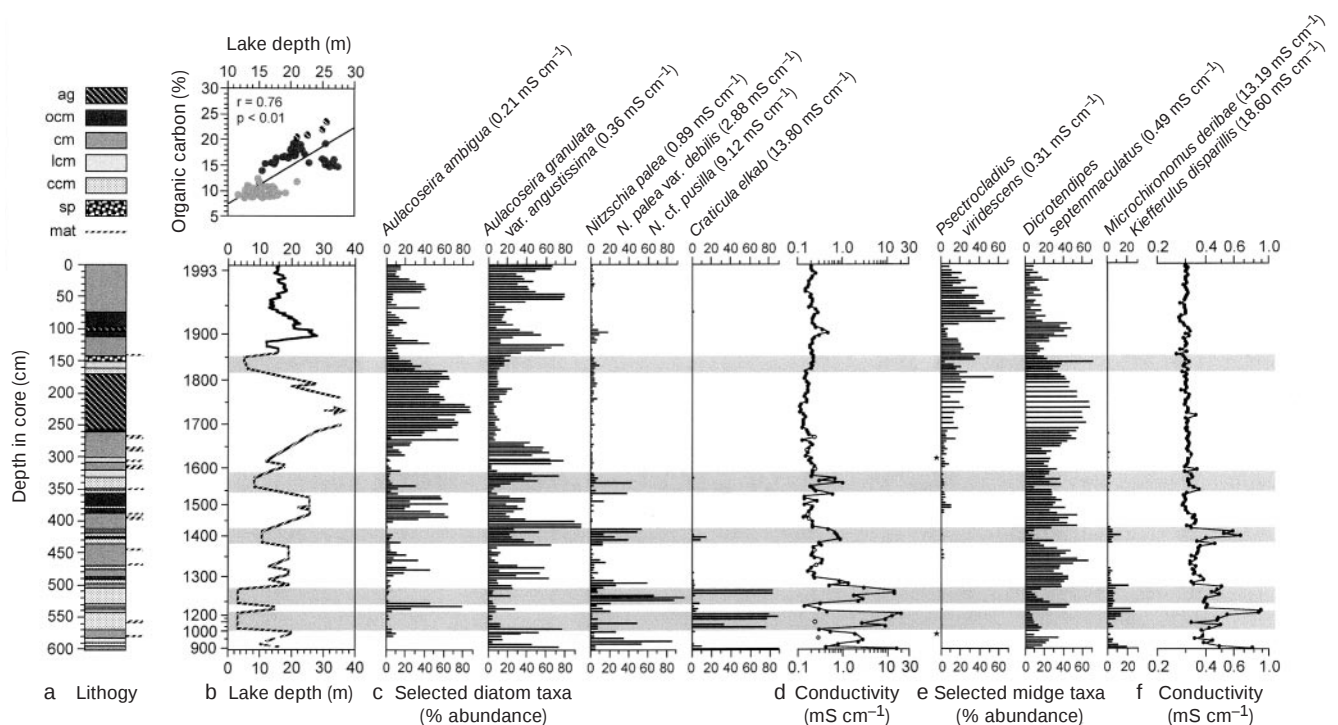


Figure 1 Sedimentological and biological evidence for past lake-depth and salinity fluctuations in the Crescent Island Crater basin of Lake Naivasha. **a**, Stratigraphic distribution of the sediment types algal gyttja (ag), organic clayey mud (ocm), clayey mud (cm), low-organic clayey mud (lcm), calcareous clayey mud (ccm), and silty peat (sp). **b**, Main panel: historical (1883–1993; solid line) and sedimentology-inferred (ref. 8, dashed line) depth variation in relation to sediment age determined by linear interpolation between dated horizons (Table 1). Top, relationship between sedimentary organic carbon and lake depth at the time of deposition in the historical record. **c**, Stratigraphic distribution of selected fossil diatom taxa in 151 horizons at 4-cm intervals (see Supplementary Information for the complete diagram); species labels include the respective salinity optima in reference lakes¹³. **d**, Diatom-inferred conductivity reconstruction; nine data points (open circles) corresponding with dense mono-specific diatom mats ('mat' in **a**) bedded in between sections of the normal sediment matrix are omitted

(< 10 yr) water-residence time⁶, but strong evaporation and highly variable river inflow cause its surface elevation to fluctuate in response to changing climate regimes. Interannual lake-level fluctuation bears the signature of rainfall anomalies linked to El Niño/Southern Oscillation (ENSO) events³, which are regionally complex. At the decadal timescale, remarkably uniform trends in the historical fluctuations of lakes throughout east Africa⁷ indicate that the spatial complexity of rainfall patterns controlled by topography and differences in maritime influence overlays a spatially coherent pattern of temporal variability; this coherent pattern is governed by large-scale atmospheric dynamics of the tropical monsoon circulation³. Since AD 1883 the depth of the CIC basin has ranged between 27 m in AD 1894 and a historic low of 12 m in 1946. The CIC basin functions as a sediment trap for a large sector of Lake Naivasha, resulting in high rates of sediment accumulation and excellent time resolution of its sediment record².

Lake sediments deposited in CIC over the past 1,100 years (Table 1) consist of six types (Fig. 1a, top) which reflect variations in bottom dynamics, clastic input, and water chemistry related to the depth of the lake and its degree of hydrologic closure at the time of deposition⁵. Analyses of modern and historical sedimentation patterns in CIC and neighbouring lakes^{5,8} revealed strong and consistent relationships between sediment composition and lake depth; these relationships can be used to reconstruct former

from the reconstruction because they may represent short-term (<<1 year) events⁹.

e, Stratigraphic distribution of selected fossil midge taxa in 149 horizons at 4-cm intervals (see Supplementary Information for the complete diagram); stars indicate missing samples, and 18 samples between 192 and 264 cm depth were combined 2-by-2 to improve representability. **f**, Chironomid-inferred conductivity reconstruction. Chironomid- and diatom-inferred conductivity estimates were calculated with weighted-averaging regression and calibration models³¹ based on salinity optima for individual taxa derived from reference data sets of bottom-mud and live diatom floras in 164 African lakes with known water chemistry¹³, and live midge faunas in 45 African lakes with chemistry data compiled from the literature¹¹. Diatom- and chironomid-based conductivity inferences are strongly correlated ($r = 0.64$, $P < 0.01$, $n = 140$), supporting the validity of both proxy salinity indicators.

fluctuations in lake depth from sediment stratigraphy⁸. For example, variation in organic-carbon content in the lithological sequence from clayey mud to gelatinous algal gyttja (Fig. 1a, top) is positively correlated with lake depth (Fig. 1b, top) because its dilution by relatively coarse-grained, inorganic littoral sediments is reduced during highstands⁸. Deposition of calcareous clayey mud reflects biogenic carbonate precipitation following the evaporative concentration of dissolved salts⁹. We interpret its stratigraphic occurrence in CIC sediments (Fig. 1a) to indicate prolonged hydrologic closure of the crater basin during lowstands below the elevation of the sill connecting it to Lake Naivasha. Deposition of silty peat with abundant remains of the aquatic plant *Ceratophyllum* is interpreted to reflect a freshwater lowstand when submerged macrophytes grew across the bottom of the crater (< 5 m depth), while the shallow basin of Lake Naivasha stood dry.

The strong documented relationship between the species distribution of diatoms and the salinity of lakes around the world has yielded regionally robust models for inferring past changes in the salinity of climate-sensitive lakes from fossil diatom assemblages¹⁰. Recent studies on the ecological distribution of aquatic midges of the family Chironomidae (Insecta: Diptera) in lakes of the Kenya Rift Valley^{11,12} suggest that their fossil remains can also provide information on past changes in salinity, as well as lake depth. Here we use a published salinity-inference model based on modern

African diatom floras¹³, and a new model based on modern African chironomid faunas¹¹, to reconstruct decade-scale fluctuations in the conductivity (a common proxy for salinity¹⁰) of CIC water during the past 1,100 years. Together, the sediment-inferred depth reconstruction (Fig. 1b) and diatom- and chironomid-inferred conductivity reconstructions (Fig. 1d,f) show that CIC experienced a long saline lowstand from about AD 1000 to 1270, interrupted by one freshwater interval in the early thirteenth century. This lowstand was followed from ~AD 1270 to 1550 by a mostly positive water balance and establishment of freshwater conditions, except for one recurrence of saline conditions dated to ~AD 1380–1420. After a second pronounced lowstand dated to ~AD 1560–1590, CIC rose to a highstand that lasted from about AD 1670 to 1770, with lake level continuously above the historical maximum reached in 1894. The diatom flora at that time (Fig. 1c; see also Supplementary Information) was dominated by *Aulacoseira ambigua*, consistent with the high light and low nutrient requirements of this species compared to *A. granulata*¹⁴, and the documented influence of lake depth on the light regime and nutrient availability in Lake Naivasha¹⁵. The chironomid fauna (Fig. 1e; see also Supplementary Information) was dominated by *Dicortendipes septemmaculatus*, consistent with the positive relationship of this species with lake depth in the historical portion of the CIC record ($r = 0.69$, $P < 0.01$, $n = 30$). This relationship results because its preferred habitat of submerged macrophytes^{11,12} expands during highstands into the large bottom areas of Lake Naivasha that slope towards the crater¹⁶. The prominent eighteenth-century highstand ended with strong lake-level decline culminating in a third lowstand during ~AD 1810–1850. In contrast to the earlier lowstands which were saline, CIC maintained freshwater conditions during this most recent drawdown, presumably because the preceding 150 years of broad confluence with the hydrologically open basin of Lake Naivasha had removed its burden of dissolved salts. Renewed lake filling in the mid-nineteenth century achieved the intermediate levels that preceded the historic highstand of the late 1880s and 1890s (Fig. 1b). Pronounced lake-level and salinity fluctuations of CIC over the

past millennium are superimposed on a long-term freshening trend that started around AD 1270; evidence for this freshening is given by the gradual disappearance of salt-tolerant diatom and chironomid taxa (Fig. 1c,e) and the expansion of stenotopic freshwater taxa¹¹ such as *Psectrocladius viridescens* (Fig. 1e).

The combined palaeoenvironmental information we derive from the CIC sediment record describes the hydrological response of Lake Naivasha to a succession of decade-scale fluctuations in the regional balance of rainfall and evaporation over the past 1,100 years. It suggests that equatorial east Africa was generally drier than today during the Mediaeval Warm Period (MWP; ~AD 1000 to 1270), and that fairly wet conditions during the Little Ice Age (LIA; ~AD 1270–1850) were interrupted around AD 1380–1420, 1560–1620 and 1760–1840 by episodes of persistent aridity more severe than any recorded drought of the twentieth century. Previously available proxy climate records from the large African Rift lakes Tanganyika¹⁷ and Turkana¹⁸ lack comparable detail, but are consistent with our inference that tropical east Africa was drier than today during the eleventh and twelfth centuries, and relatively wet from the late thirteenth century to the mid-eighteenth century. Sediment records from shallow lakes in the region are typically incomplete, but a LIA-period highstand of Lake Chilwa in Malawi dated to AD 1650–1760¹⁹ is coeval with the most prominent highstand of Lake Naivasha. The inferred prevalence of drought in tropical Africa during the MWP, and the highest rainfall broadly coincident with the lowest LIA temperatures in mid-latitude regions of the Northern Hemisphere²⁰, implies that the latitudinal pattern of century-scale climate anomalies during the past 1,100 years was opposite to that which occurred on millennial timescales during the last glaciation²¹, the Younger Dryas²², and early Holocene²³.

Comparison of the Naivasha record with that of reconstructed atmospheric ¹⁴CO₂ production, a proposed proxy for long-term changes in solar radiation²⁴, reveals that the MWP period of inferred African aridity and all three severe drought events of the past 700 years were broadly coeval with phases of high solar radiation, and intervening periods of increased moisture were coeval with phases

Table 1 Chronology of the sediment record from Crescent Island Crater basin

Core depth (cm)	Dated horizon or material	²¹⁰ Pb or ¹⁴ C (years AD or BP)	Calendar year AD (1σ range)
Sediment age*			
0	Sediment surface		1993.6
10–11	<i>Salvinia</i> remains		1989
16–17	<i>Salvinia</i> remains		1982
30–31	<i>Salvinia</i> remains		1969
Age of stratigraphic marker horizons†			
49–50	<i>Daphnia</i> resting eggs	1958 ± 4 AD	1958 (1954–1962)
73–74	Transition ocm to cm§	1940 ± 8 AD	1940 (1932–1948)
113–114	Transition cm to ocm§	1890 ± 35 AD	1890 (1855–1925)
Radiocarbon dates‡			
150–178	Grass charcoal	120 ± 50 BP	1819, 1860, 1917 (1803–1937) 1718, 1702 (1680–1753)
350–374	<i>Acacia</i> leaflets, grass charcoal	320 ± 40 BP	1530, 1537, 1635 (1511–1647)
374–398	Grass charcoal	480 ± 50 BP	1436 (1414–1449)
419–420	Sedge rhizome	580 ± 40 BP	1400 (1314–1411)
486–506	<i>Acacia</i> leaflets, grass charcoal	770 ± 50 BP	1278 (1229–1288)
506–510	<i>Acacia</i> twig	(930 ± 50 BP)	1052, 1085, 1121, 1139, 1156 (1028–1178)
538–542	Charred wood	820 ± 50 BP	1229 (1191–1278)
542–546	Wood	(1100 ± 60 BP)	982 (938–1000) 910, 906 (892–930)
558–562	Unburned grass stem	(1250 ± 50 BP)	779 (689–873)
562–566	<i>Acacia</i> wood, leaflet	1030 ± 50 BP	1014 (983–1029)
562–566	Sedge rhizome	1040 ± 50 BP	1011 (978–1026)
592–602	Grass charcoal	1140 ± 60 BP	890 (869–985)

* Sediment age at depth based on fossil evidence of three documented outbreaks of the exotic water fern *Salvinia molesta*.⁵

† Age of stratigraphic marker horizons (years AD ± 1σ) determined by correlation with the equivalent horizons in a ²¹⁰Pb-dated sediment core from the main basin of Lake Naivasha.⁵

‡ Radiocarbon dates (years BP ± 1σ) determined by accelerator mass spectrometry at the Lawrence Livermore National Laboratory. Agreement between the ages of sedge rhizome and *Acacia* wood retrieved from depth interval 562–566 cm suggests that ¹⁴C dates obtained on emergent aquatic vegetation and terrestrial plant material can be used in combination. Radiocarbon dates were converted to calendar year using the bidecadal calibration curve²⁴, and the continuous sediment chronology was constructed by linear interpolation between these calendar ages. Three out-of-sequence ¹⁴C ages (in parentheses) on single macrofossils recovered from calcareous clayey mud between 506 and 572 cm depth are omitted from the chronology on the basis that they may be derived from older deposits reworked during lowstands⁵. In cases where conversion yielded non-unique calendar ages, a fourth-order polynomial regression of all retained dates was used to determine the relative probability of alternative calendar ages based on minimization of downcore variation in sedimentation rate.

§ Sediment type codes as in Fig. 1.

of low solar radiation (Fig. 2). Hence, variation in solar radiative output may have contributed to decade-scale rainfall variability in equatorial east Africa. Specifically, highest inferred rainfall of the past 1,100 years coincided with the 'Maunder minimum' of solar radiation (AD 1645–1715; ref. 25). The Naivasha record suggests that late-sixteenth-century drought partly preceded peak solar radiation, but the time difference involved is within the uncertainty range of the supporting radiocarbon chronology (Table 1). Similar to historical lake-level records in the region⁷, our 1,100-year reconstruction reflects lake response to decade-scale clusters of years with above- or below-average rainfall, and can thus be considered to reflect the long-term history of water-resource availability in continental east Africa. The amplifying effect of river inflow²⁶ from a large Rift Valley drainage basin allowed the possibly subtle effect of solar forcing on African rainfall to be manifested in Naivasha's lake-level record.

Comparison of this record with the pre-colonial history of east Africa recounted in oral traditions⁴ testifies to the importance of rainfall and drought in agricultural and pastoral societies (Fig. 2). In the six centuries before AD 1895, evidence for drought-induced famine, political unrest, and large-scale migration of indigenous peoples is concentrated in three periods around AD 1390–1420 (Wamara), AD 1560–1625 (Nyarubanga), and AD 1760–1840 (Lapanarat-Mahlatule) that match the reconstructed sequence of Lake Naivasha lowstands (Fig. 2). The intervening 'first and second Age of Prosperity' (ref. 27)—relatively uneventful periods of political stability, consolidation of kingdoms, agricultural success and population growth—are coeval with the two most prominent Lake Naivasha highstands. In the near-complete absence of archeological data²⁸, the Lake Naivasha record establishes the environmental background for political and cultural change in pre-colonial east Africa, indicates strong links between cultural development and

climate change, and reinforces the chronology of historical events as derived from oral traditions.

Lack of long palaeomoisture records from low-latitude regions has thus far limited our understanding of global climate variability at decade-to-century timescales¹, and hindered efforts to identify physical feedback mechanisms that may strengthen the direct temperature effect of long-term changes in solar radiation²⁹. Our results corroborate findings from north-temperate dryland regions³⁰ that instrumental climate records are inadequate to appreciate the full range of natural variation in drought intensity at timescales relevant to socio-economic activity. The magnitude of natural decade-scale rainfall variability in sub-humid east Africa implies that sustainable development and protection of food security will require agricultural management strategies adjusted to major long-term variation in water-resource availability, irrespective of any future effects of anthropogenic climate change on the hydrological cycle. □

Received 8 April; accepted 16 November 1999.

- Houghton, J. T. et al. (eds). *Climate Change 1995 The Science of Climate Change* (Cambridge Univ. Press, 1995).
- Nicholson, S. E. *The Physical Geography of Africa* (eds Goudie, A. S., Adams, W. M. & Orme A.) 60–75 (Oxford Univ. Press, 1995).
- Nicholson, S. E. In *The Limnology, Climatology and Paleoclimatology of the East African Lakes* (eds Johnson, T. C. & Odada, E.) 25–56 (Gordon & Breach, Newark, 1996).
- Webster, J. B. In *Chronology, Migration and Drought in Interlacustrine Africa* (ed. Webster J. B.) 1–37 (Longman & Dalhousie Univ. Press, Dalhousie, 1979).
- Verschuren, D. Sedimentation controls on the preservation and time resolution of climate-proxy records from fluctuating lakes. *Quat. Sci. Rev.* **18**, 821–837 (1999).
- Ojiambo, B. S. & Lyons, W. B. In *The Limnology, Climatology and Paleoclimatology of the East African Lakes* (eds Johnson, T. C. & Odada, E.) 267–278 (Gordon & Breach, Newark, 1996).
- Nicholson, S. E. In *Environmental Change and Response in East African Lakes* (ed. Lehman J. T.) 7–35 (Kluwer, Dordrecht, 1998).
- Verschuren, D. Reconstructing fluctuations of a shallow East African lake during the past 1800 years from sediment stratigraphy in a submerged crater basin. *J. Paleolimnol.* (in the press).
- Kilham, P. Mechanisms controlling the chemical composition of lakes and rivers: data from Africa. *Limnol. Oceanogr.* **35**, 80–83 (1990).
- Fritz, S. C., Cumming, B. F., Gasse, F. & Laird, K. R. In *Diatoms: Applications to the Environmental and Earth Sciences* (eds Stoermer, E. & Smol, J. P.) 41–72 (Cambridge Univ. Press, 1999).
- Verschuren, D. Taxonomy and ecology of fossil Chironomidae (Insecta, Diptera) from Rift Valley lakes in central Kenya. *Arch. Hydrobiol. Suppl.* **107**, 467–512 (1997).
- Verschuren, D., Tibby, J., Sabbe, K. & Roberts, N. Effects of lake level, salinity and substrate on the benthic invertebrate community of a shallow tropical-African lake ecosystem. *Ecology* (in the press).
- Gasse, F., Juggins, S. & Ben Khelifa, L. Diatom-based transfer functions for inferring hydrochemical characteristics of African paleolakes. *Paleogeogr. Paleoclimatol. Paleoecon.* **117**, 31–54 (1995).
- Kilham, P., Kilham, S. S. & Hecky, R. E. Hypothesized resource relationships among African planktonic diatoms. *Limnol. Oceanogr.* **31**, 1169–1181 (1986).
- Harper, D. M., Phillips, G., Chilvers, A., Kitaka, N. & Mavuti, K. Eutrophication prognosis for Lake Naivasha, Kenya. *Verh. Int. Verein. Limnol.* **25**, 861–865 (1993).
- Harper, D. M. The ecological relationships of aquatic plants at Lake Naivasha, Kenya. *Hydrobiologia* **232**, 65–71 (1992).
- Cohen, A. S., Talbot, M. R., Awramik, S. M., Dettman, D. L. & Abell, P. Lake level and paleoenvironmental history of Lake Tanganyika, Africa, as inferred from late Holocene and modern stromatolites. *Geol. Soc. Am. Bull.* **109**, 444–460 (1997).
- Halfman, J. D., Johnson, T. C. & Finney, B. P. New AMS dates, stratigraphic correlations and decadal climatic cycles for the past 4 ka at Lake Turkana, Kenya. *Paleogeogr. Paleoclimatol. Paleoecon.* **111**, 83–98 (1994).
- Owen, R. B. & Crossley, R. Recent sedimentation in lakes Chila and Chiuta, Malawi. *Paleoecol. Afr.* **20**, 109–117 (1989).
- Mann, M. E., Bradley, R. S. & Hughes, M. K. Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. *Geophys. Res. Lett.* **26**, 759–762 (1999).
- Kutzbach, J. E. & Street-Perrott, F. A. Milankovitch forcing of fluctuations in the level of tropical lakes from 18 to 0 kyr BP. *Nature* **317**, 130–134 (1985).
- Roberts, C. N. et al. Timing of the Younger Dryas event in East Africa from lake-level changes. *Nature* **366**, 146–148 (1993).
- Street-Perrott, F. A. & Perrott, R. A. Abrupt climate fluctuations in the tropics: the influence of Atlantic Ocean circulation. *Nature* **348**, 607–612 (1990).
- Stuiver, M. & Reimer, P. J. Extended ¹⁴C data base and revised CALIB 3.0 ¹⁴C age calibration program. *Radiocarbon* **35**, 215–230 (1993).
- Lean, J., Beer, J. & Bradley, R. S. Reconstruction of solar irradiance since 1610: implications for climate change. *Geophys. Res. Lett.* **22**, 3195–3198 (1995).
- Knox, J. C. Large increases in flood magnitude in response to modest changes in climate. *Nature* **361**, 430–432 (1993).
- Webster, J. B. Drought, migration and chronology in the Lake Malawi littoral. *Transafr. J. Hist.* **9**, 70–90 (1980).
- Connah, G. *African Civilizations* (Cambridge Univ. Press, 1987).
- Rind, D. & Overpeck, J. Hypothesized causes of decade-to-century-scale climate variability: climate model results. *Quat. Sci. Rev.* **12**, 357–374 (1993).
- Woodhouse, C. A. & Overpeck, J. T. 2000 years of drought variability in the central United States. *Bull. Am. Meteorol. Soc.* **79**, 2693–2714 (1998).

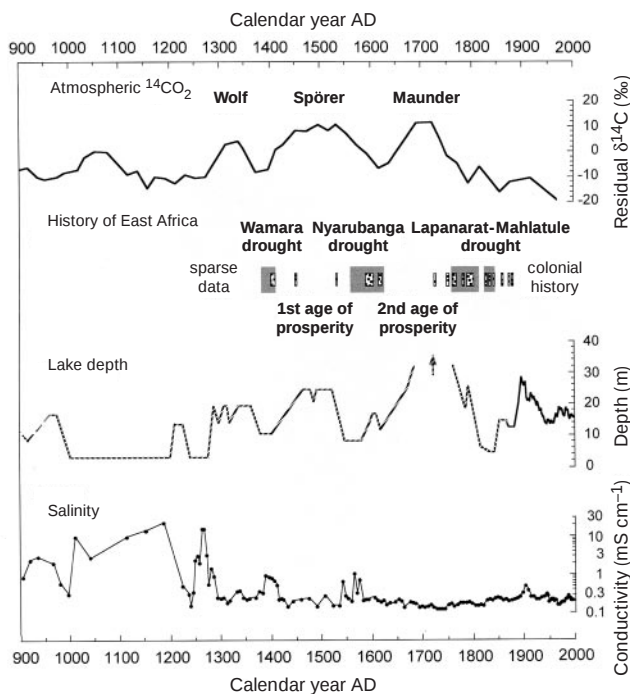


Figure 2 Comparison of Crescent Island Crater history with documented and reconstructed climate-proxy data. Shown is a comparison of reconstructions of the depth and salinity of Crescent Island Crater lake with the decadal record of atmospheric ¹⁴CO₂ production²⁴ as a proxy for solar radiation, and the pre-colonial history of east Africa. Grey bars indicate evidence of drought-related political upheaval recorded in oral tradition, genealogically dated using a 27-yr dynastic generation⁴. Stippled bars compile the evidence of severe drought events from various archival records² including the (incomplete) record of Nile River discharge.

31. Line, J. M., ter Braak, C. J. F. & Birks, H. J. B. WACALIB version 3. 3: A computer program to reconstruct environmental variables from fossil assemblages by weighted averaging and to derive sample-specific errors of prediction. *J. Paleolimnol.* **10**, 147–152 (1994).

Supplementary information is available on Nature's World-Wide web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank K. Mavuti, B. Ammann, F. Janssen and H. E. Wright for field assistance; the Lake Naivasha Riparian Association for lake access; F. Gasse for use of the African diatom reference and calibration data sets; and S. Fritz, F. Gasse, K. Kelts, F. Oldfield, J. P. Smol, B. Tinsley and H. E. Wright for comments on the manuscript. This work was supported by NSF, the Quaternary Paleocology Program at the University of Minnesota, and by fellowships from NOAA and FWO-Vlaanderen (D.V.). The fieldwork was conducted with permission from the Office of the President of the Republic of Kenya to K. Mavuti.

Correspondence should be addressed to D.V. (e-mail: dirk.verschuren@rug.ac.be). The data presented here are archived at the World Data Center-A for Paleoclimatology.

Precise climate monitoring using complementary satellite data sets

Frank J. Wentz & Matthias Schabel

Remote Sensing Systems, 438 First Street, Suite 200, Santa Rosa, California 95401, USA

Observations from Earth-orbiting satellites have been a key component in monitoring climate change for the past two decades. This has become possible with the availability of air temperatures from the Microwave Sounding Unit (MSU)¹ since 1979, sea surface temperatures from the Advanced Very High Resolution Radiometer (AVHRR)² since 1982 and, most recently, measurements of atmospheric water vapour content from the Special Sensor Microwave Imager (SSM/I)³ since 1987. Here we present a detailed comparison of each pair of these three time series, focusing on both interannual and decadal variations in climate. We find a strong association between sea surface temperature, lower-tropospheric air temperature and total column water-vapour content over large oceanic regions on both time scales. This lends observational support to the idea of a constant relative humidity model having a moist adiabatic lapse rate. On the decadal timescale, the combination of data sets shows a consistent warming and moistening trend of the marine atmosphere for 1987–1998.

The detection and measurement of small changes in the Earth's climate require extremely precise global observations of a broad spectrum of complementary physical variables. In this endeavour, satellite observations are playing an increasingly important role. As compared to conventional *in situ* observations, satellites provide daily near-global coverage with a very high statistical precision that results from averaging millions of individual observations. Here we present a three-way intercomparison of three satellite-derived climate variables that are closely correlated in the marine boundary layer and lower troposphere: sea surface temperature T_s , lower-tropospheric air temperature T_A , and vertically integrated, or columnar, atmospheric water vapour W . This intercomparison allows us to determine the coupling of the three variables on interannual and decadal timescales and investigate the accuracy of the individual time series^{4,5,6}.

The T_A and T_s data sets are standard products that have been available for some time^{1,2}, whereas the W data set is a relatively new product³ beginning in 1987 with the launch of the special sensor microwave imager (SSM/I), a multichannel microwave radiometer.

Since 1987 four more SSM/Is have been launched, providing an uninterrupted 12-year time series. Imaging radiometers before SSM/I were poorly calibrated, and as a result early water-vapour studies⁷ were unable to address climate variability on interannual and decadal timescales.

To retrieve water vapour from the SSM/I observations, we use a physically based algorithm that simultaneously computes water vapour, wind speed, and cloud water by directly matching the observations to a radiative transfer model³. The primary channel for W is at 22 GHz, which is centred on a water-vapour absorption line. Additional dual-polarization channels at 19 and 37 GHz are used to remove crosstalk (that is, a signal from one variable aliasing into another) attributable to wind and clouds. A monthly climatology is used to specify the regional and seasonal variations of T_s . The T_A and T_s decadal time series are used to remove a small crosstalk component on interannual timescales. This correction for interannual crosstalk is quite small, reducing the amplitude of the interannual variability in W by about 8%. Other water-vapour retrieval algorithms have been developed, but most do not fully account for crosstalk or do not use the 22-GHz channel^{8,9}, thereby limiting their ability to measure the small interannual climate signal. The retrieval of W requires the radiometrically cold background of the sea surface and cannot be done over land, which is highly emissive.

For the intercomparisons, we use Reynolds' 1982–1998 optimally interpolated sea surface temperature data set² and the 1979–1998 version-c1 MSU lower-tropospheric air temperature¹⁰ with a correction for orbital decay that is added to the zonal averages⁶. (We do not use the new version-d MSU product because it has not yet been published. However, for the intercomparison period (1987–1998) over the oceans, there is little difference (0.01 K per decade) between version c1 plus orbit decay and version d.) The three variables are geographically and temporally sampled in the same manner to ensure an unbiased comparison. Figure 1 shows the deseasonalized anomalies of W , T_s , and T_A . Results are shown for three zones: the northern extratropics (20°N–60°N), the tropics (20°N–20°S), and the southern extratropics (20°S–60°S). The seasonal cycle has been removed and the residuals have been low-pass filtered by convolution with a gaussian distribution having a ± 90 -day width at half-peak value. The variation of water vapour with latitude is large (5–60 mm), so it is more convenient to express W in terms of a percentage change. To compute percentages, the daily zonal W is divided by the 12-year mean value for the zone. The prominent El Niño/Southern Oscillation signal is clearly evident in the tropics, peaking in January 1983, December 1987 and December 1997. For the 1983 and 1997 El Niños, the T_s signal leads T_A by a few months, but the 1987 El Niño shows no such lag. The scaling in Fig. 1 is based on radiosonde observations of mid-latitude seasonal variability, which show $\Delta W/\Delta T_s = 11\% \text{ K}^{-1}$ and $\Delta T_A/\Delta T_s = 1.6$. To display this scaling, the T_A time series have been divided by 1.6, and the vapour scale on the right side of the figure is chosen so that $11\% = 1 \text{ K}$.

Three statistics for the 1987–1998 period are given in Table 1: the correlation coefficient, the decadal trend, and the scaling coefficient, which is the ratio of standard deviations of the time series. The three variables are closely coupled, as has been shown by previous studies on shorter timescales of 2–5 years (refs 9 and 11). In the tropics, where the El Niño/Southern Oscillation cycle dominates the interannual variability, the correlation coefficients are very high: W with T_s is 0.98, W with T_A is 0.93, and T_s with T_A is 0.94. The corresponding tropical scaling coefficients are $\Delta W/\Delta T_s = 9.2\% \text{ K}^{-1}$, $\Delta T_A/\Delta T_s = 1.4$ and $\Delta W/\Delta T_A = 6.7\% \text{ K}^{-1}$, which are similar to the mid-latitude radiosonde seasonal scalings. These tropical scaling relationships are consistent with a constant relative humidity (H_{rel}) model having a moist adiabatic lapse rate (MALR)¹². The Clausius–Clapeyron equation for saturation vapour pressure¹³ predicts that the water-vapour density, and hence columnar content, increases

with air temperature if H_{rel} stays constant. The rate of increase varies from 6.0% to 7.5%, depending on air temperature, with a value near 6.7% K^{-1} for the tropics. Other satellite studies^{7,8,11} on the seasonal variability of W , when averaged zonally, also support the constant- H_{rel} assumption.

According to the MALR model, T_A variability is greater than T_S variability, as is shown by the satellite observations. For example, if T_S is 293K (294K), MALR gives a temperature of 276.4 K (277.8) at 3.5 km. Thus a change of 1K at the surface corresponds to a change of 1.4 K at an altitude of 3.5 km, which is the nominal altitude for the MSU T_A . We are assuming the surface air temperature equals T_S , which is generally the case over the ocean to within a few tenths of a degree. Although significant differences have been observed between the MALR and the actual lapse rate for some environments such as over cold land surfaces¹², the MALR is in good agreement with *in situ* observations over the ocean. We compared the MALR averaged over the lower 3.5 km of the troposphere with radiosonde observations from 56 small islands around the world during 1987–1991 and found a mean and r.m.s. difference of -0.2 K km^{-1} and 1.0 K km^{-1} , respectively.

Our results for large-scale zonal averages differ from radiosonde regional studies¹⁴, which show a relatively weak W versus T_A correlation on interannual timescales. The weak correlation is probably due to regional variability in atmospheric stratification and convection. It appears that the zonal averaging effectively cancels the regional variability and produces a result that is well represented by the relatively simple constant- H_{rel} MALR model. However, this model is only applicable for large zonal averages, and not necessarily to regional processes.

We now examine the decadal trends for the 1987–1998 period. Long-term trends are a difficult statistic to derive from satellite data because of problems with satellite intercalibration and sensor drift^{4,5,6,10}. However, such trends are important because they may be indicators of anthropogenic climate change. Although a 12-year period is relatively short for inferring decadal trends, the time

period does give us the opportunity to compare three independent data sets. As the measurement errors in the three records are largely independent, the three time series, in combination, may be capable of defining the climate trends with greater confidence than any one of them by itself.

In the tropics the trends are 2.1% per decade, 0.23 K per decade, and 0.22 K per decade for W , T_S and T_A , whereas the global trends are 1.8% per decade, 0.15 K per decade, and 0.10 K per decade, respectively. The ratio of the W and T_S trends closely matches the seasonal and interannual scaling coefficient. However, the ratio of the T_A and T_S trends ranges between 0.7 and 1, rather than the 1.4 scaling exhibited by the interannual variability. It is not clear if this discrepancy is real or due to small sensor-calibration problems. The overlap interval between the T_S and T_A time series extends back to 1982, which allows for the comparison of these two variables on a longer timescale. The tropical and global 1982–1998 decadal trends for T_S are 0.13 K per decade and 0.12 K per decade and for T_A are 0.20 K per decade and 0.19 K per decade respectively, revealing a T_A/T_S scaling consistent with the seasonal and interannual scaling. On the longest timescale of 1979–1998, for which only the MSU data are available, the tropical and global T_A trends are 0.12 K per decade and 0.13 K per decade, respectively.

Regional radiosonde studies have also indicated a warming and moistening of the atmosphere over the last 10–20 years. Vapour trends over North America from 1973–1993 were reported to be 3–7% per decade¹⁵ and to be 1% per decade over China from 1970–1990 (ref. 16). *In situ* observations also show a general warming and moistening over the tropical Pacific from the early 1970s to the early 1990s (refs 14 and 17). An apparent contradiction to these results is a study involving satellite vapour retrievals from infrared observations that reports widespread drying in the tropics from 1979–1995 (ref. 18). However, inferring vapour trends from infrared measurements is difficult, requiring continuous calibration to radiosondes, and it appears that the reported drying is probably spurious, caused by problems with the radiosondes used for calibration¹⁹.

The decadal trend for W comes from a least-squares fit to deseasonalized zonal daily averages. There are two ways to interpret the statistical significance of this decadal trend. First, if interannual variability is considered to be a source of noise masking the linear trend, then the time series can be modelled as a first-order autoregression²⁰. By computing the lag-1 correlation coefficient and the residual variance, we find that the standard error in the trend estimate is 0.5% per decade for the tropics and 0.3% per decade globally. Second, the trend may be interpreted as a first-order statistic of the interannual variability, with the only source of error being the satellites' ability to measure the true daily zonal averaged W . This satellite measurement error, which results from retrieval error and incomplete spatial/temporal sampling, is estimated by comparing simultaneous observations from different satellites. We find that the standard error in the daily W is 1.5%

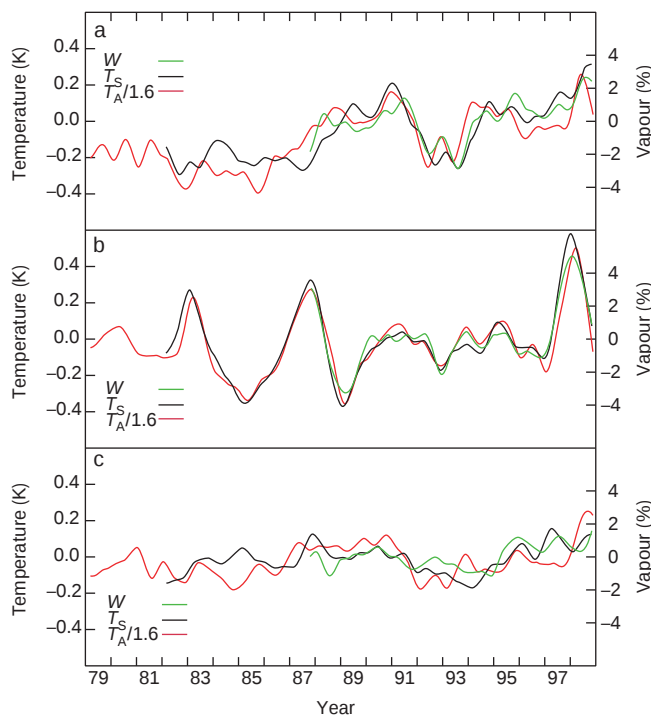


Figure 1 Anomaly time series of three climate variables. **a, b, c**, Water vapour W (green), sea surface temperature T_S (black), and air temperature T_A (red) are shown for the northern extratropics (20°N–60°N; **a**), tropics (20°S–20°N; **b**), and southern extratropics (20°S–60°S; **c**). Air temperature is divided by 1.6.

Table 1 Statistics on interannual and decadal variability

	20°N–60°N	20°S–20°N	20°S–60°S
Decadal trend			
W (% per decade)	1.9	2.1	1.0
T_S (K per decade)	0.20	0.23	0.03
T_A (K per decade)	0.04	0.22	0.00
Correlation			
W with T_S	0.84	0.98	0.73
W with T_A	0.71	0.93	0.36
T_A with T_S	0.69	0.94	0.56
Scaling coefficient			
W/T_S (% K^{-1})	8.5	9.2	8.6
W/T_A (% K^{-1})	6.8	6.7	4.6
T_A/T_S	1.2	1.4	1.9

and is uncorrelated from one day to the next. This gives a standard error in the trend estimate of 0.08% per decade²⁰. An additional error is the uncertainty in determining the intersatellite biases, which are estimated from data collected during periods of satellite overlap (1–4 years). Monte Carlo simulations show that the uncertainty in specifying the intersatellite biases introduces a 0.2% per decade standard error in the trend estimate. These estimates of trend error do not include the additional uncertainty involved with questions of prediction or representation (that is, how well the 1987–1998 period represents longer-term climate change).

As a check on the error analysis, we divide the SSM/I data into two sets: local morning and local evening. For each set, completely independent derivations of the decadal trends for the three latitude zones give a maximum difference between the morning and evening trends of 0.2% per decade. This agreement indicates that diurnal effects only weakly influence the SSM/I data, in marked contrast to the significant diurnal problems affecting the MSUs. A complicated procedure is required to remove the diurnal effect from the MSU observations¹⁰, and this possibly introduces spurious trends.

The retrieval of W from SSM/I observations is less problematic than the T_S and T_A retrievals in other ways as well. The 22-GHz radiance observed by SSM/I is directly related to W , whereas the MSU measures the middle-to-upper troposphere and then infers the lower-tropospheric temperature²¹. Furthermore, assuming a constant H_{rel} , the 22-GHz water-vapour radiance is three times more sensitive to changes in air temperature than MSU 54-GHz radiance. For AVHRR, a continuous calibration with *in situ* data is required to remove the cooling effect of atmospheric aerosols such as those emitted by volcanic eruptions²² and to correct for instrument drift and intersatellite biases². The robustness and accuracy of the SSM/I W retrieval makes it a useful validation tool for the more complex retrievals of T_A and T_S , although we must keep in mind that they are different physical variables.

The consistency we now see emerging from various satellite observations provides new information on climate dynamics and should help resolve some of the past controversies concerning errors in the satellite data. It is notable that a relatively simple constant- H_{rel} plus MALR model closely predicts the observed interannual and decadal variations of W , T_S and T_A when zonally averaged over the tropics. Furthermore, the evidence here shows that the marine atmosphere has significantly warmed and moistened over the past decade.

Received 8 March; accepted 26 November 1999.

- Spencer, R. W. & Christy, J. R. Precise monitoring of global temperature trends from satellites. *Science* **247**, 1558–1562 (1990).
- Reynolds, R. W. & Smith, T. M. Improved global sea surface temperature analyses using optimum interpolation. *J. Clim.* **7**, 929–948 (1994).
- Wentz, F. J. A well-calibrated ocean algorithm for special sensor microwave/imager. *J. Geophys. Res.* **102**, 8703–8718 (1997).
- Hurrell, J. W. & Trenberth, K. E. Spurious trends in satellite MSU temperatures from merging different satellite records. *Nature* **386**, 164–167 (1997).
- Hurrell, J. W. & Trenberth, K. E. Difficulties in obtaining reliable temperature trends: reconciling the surface and satellite microwave sounding unit records. *J. Clim.* **11**, 945–967 (1998).
- Wentz, F. J. & Schabel, M. Effects of satellite orbital decay on MSU lower tropospheric temperature trends. *Nature* **394**, 661–664 (1998).
- Stephens, G. L. On the relationship between water vapor over the oceans and sea surface temperature. *J. Clim.* **3**, 634–645 (1990).
- Jackson, D. L. & Stephens, G. L. A study of SSM/I-derived columnar water vapor over the global oceans. *J. Clim.* **8**, 2025–2038 (1995).
- Randel, D. L. *et al.* A new global water vapor dataset. *Bull. Am. Meteorol. Soc.* **77**, 1233–1246 (1996).
- Christy, J. R., Spencer, R. W. & Lobl, E. S. Analysis of the merging procedure for the MSU daily temperature time series. *J. Clim.* **11**, 2016–2041 (1998).
- Bony, S., Duvel, J. -P. & Le Treut, H. Observed dependence of the water vapor and clear-sky greenhouse effect on sea surface temperature: comparison with climate warming experiments. *Clim. Dyn.* **11**, 307–320 (1995).
- Stone, S. H. & Carlson, J. H. Atmospheric lapse rate regimes and their parameterization. *J. Atmos. Sci.* **36**, 415–423 (1979).
- Peixoto, J. P. & Oort, A. H. *Physics of Climate* (American Institute of Physics Press, New York, 1991).
- Gaffen, D. J., Elliott, W. P. & Robock, A. Relationships between tropospheric water vapor and surface temperature as observed by radiosondes. *Geophys. Res. Lett.* **19**, 1839–1842 (1992).
- Ross, R. J. & Elliott, W. P. Tropospheric water vapor climatology and trends over north America: 1973–93. *J. Clim.* **9**, 3561–3574 (1996).

- Zhai, P. & Eskridge, R. E. Atmospheric water vapor over China. *J. Clim.* **10**, 2643–2652 (1997).
- Gutzler, D. Low-frequency ocean–atmosphere variability across the tropical western Pacific. *J. Atmos. Sci.* **53**, 2773–2785 (1996).
- Schroeder, S. R. & McGuirk, J. P. Widespread tropical atmospheric drying from 1979 to 1995. *Geophys. Res. Lett.* **25**–9, 1301–1304 (1998).
- Ross, R. J. & Gaffen, D. J. Comment on Widespread tropical atmospheric drying from 1979 to 1995, by Schroeder and McGuirk. *Geophys. Res. Lett.* **25**–23, 4357–4358 (1998).
- Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* (Academic, New York, 1995).
- Spencer, R. W. & Christy, J. R. Precision and radiosonde validation of satellite gridpoint temperature anomalies. Part II: A tropospheric retrieval and trends during 1979–1990. *J. Clim.* **53**, 858–866 (1992).
- Reynolds, R. W. Impact of Mount Pinatubo aerosols on satellite-derived sea surface temperatures. *J. Clim.* **6**, 768–775 (1993).

Acknowledgements

This work was supported by NASA as part of their pathfinder Data Set program.

Correspondence and requests for materials should be addressed to F.J.W. (e-mail: wentz@remss.com).

The Pleistocene serpent *Wonambi* and the early evolution of snakes

John D. Scanlon*† & Michael S. Y. Lee*

* Department of Zoology, University of Queensland, Brisbane QLD 4072, Australia

† Department of Biological Sciences, University of New South Wales, Sydney, NSW 2052, Australia

The Madtsoiidae were medium sized to gigantic snakes with a fossil record extending from the mid-Cretaceous to the Pleistocene, and spanning Europe, Africa, Madagascar, South America and Australia^{1–3}. This widely distributed group survived for about 90 million years (70% of known ophidian history), and potentially provides important insights into the origin and early evolution of snakes. However, madtsoiids are known mostly from their vertebrae, and their skull morphology and phylogenetic affinities have been enigmatic. Here we report new Australian material of *Wonambi*, one of the last-surviving madtsoiids^{4–6}, that allows the first detailed assessment of madtsoiid cranial anatomy and relationships. Despite its recent age, which could have overlapped with human history in Australia, *Wonambi* is one of the most primitive snakes known—as basal as the Cretaceous forms *Pachyrhachis*⁷ and *Dinilysia*⁸. None of these three primitive snake lineages shows features associated with burrowing, nor do any of the nearest lizard relatives of snakes (varanoids). These phylogenetic conclusions contradict the widely held ‘subterranean’ theory of snake origins^{9–12}, and instead imply that burrowing snakes (scolecophidians and anilioids) acquired their fossorial adaptations after the evolution of the snake body form and jaw apparatus in a large aquatic or (surface-active) terrestrial ancestor.

Serpentes Linnaeus, 1758
Madtsoiidae Hoffstetter, 1961
Wonambi Smith, 1976

Revised diagnosis. Neural spines of vertebrae high, sloping, posterodorsally, with sharp-edged anterior lamina extending to near anterior edge of zygosphenes; transverse processes extending laterally beyond zygapophyses in most trunk vertebrae, synapophyses with concave dorsal edge in lateral view; zygosphenes relatively narrow, with steep facets (20–30° from vertical); zygapophyses inclined 20° or more above horizontal; haemal keel in middle and posterior trunk region narrow and weakly defined laterally, but often distinctly bifid or trifid on the posterior third of the centrum.

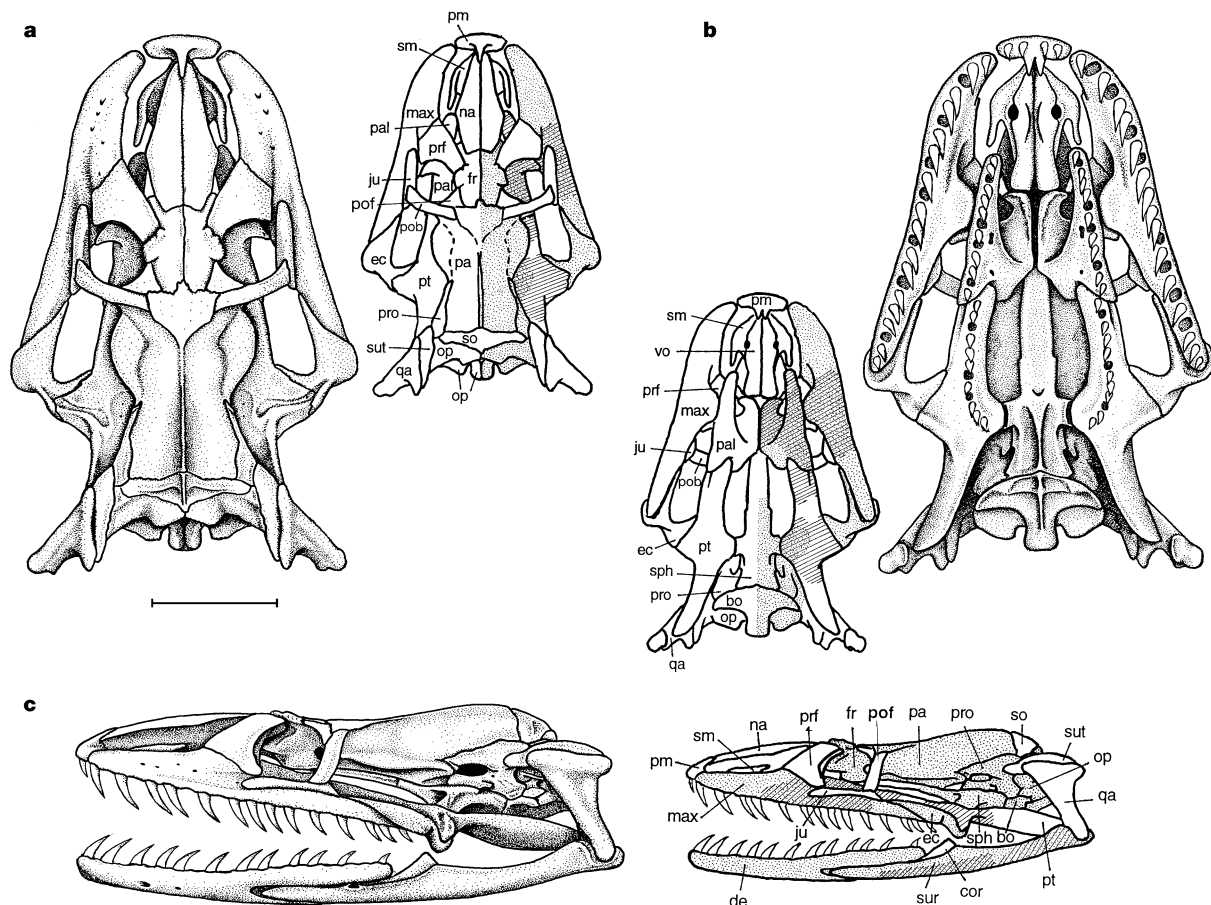


Figure 1 Reconstruction of the skull of *Wonambi naracoortensis* Smith, 1976. **a**, Dorsal, **b**, palatal and **c**, lateral views. Small outline drawings indicate parts known in *W. naracoortensis* (stipple) and in *W. barriei* (cross-hatching). Scale bar, 30 mm (based on the largest known individual at type locality, FU1762). bo, basioccipital; ec, ectopterygoid;

cor, coronoid; de, dentary; fr, frontal; ju, jugal; max, maxilla; na, nasal; op, opisthotic; pal, palatine; pa, parietal; pm, premaxilla; prf, prefrontal; pob, postorbital; pof, postfrontal; pro, prootic; pt, pterygoid; qa, quadrate; sm, septomaxilla; so, supraoccipital; sph, sphenoid; sur, surangular; sut, supratemporal; vo, vomer.

Pterygoid tooth row near middle of bone, away from medial edge, and basiptyergoid facet narrow and facing medially as much as dorsally; ectopterygoid process of pterygoid triangular in palatal view. Maxilla and dentary relatively elongate and depressed; maxilla with deep, anterolaterally directed trough on suborbital surface.

Wonambi naracoortensis Smith, 1976

Diagnosis. Large size (trunk vertebrae more than 20 mm and often greater than 30 mm wide, mandible length up to 160 mm, total length estimated to exceed 5 m); relatively small neural canal and deep zygosphenoid; zygapophyses inclined about 25° above horizontal; parazygantral foramina often small and multiple, weakly distinguished from pits in the same region. Palatine broad (mediolateral dimension of choanal process greater than that of tooth-bearing portion of palatine).

Material. Abundant vertebrae (including holotype mid-trunk vertebra, South Australian Museum P16168), ribs, and the following elements of the skull: complete maxillae, prootic, exoccipital-opisthotics, basioccipital, dentaries, compound mandibular elements; near-complete sphenoid, frontal, palatine, ectopterygoid, and parietal; and fragment of pterygoid. Most of these are represented in the partial skeletons from Henschke's Cave, Naracoorte (SAM P30178)⁵, but some have not been identified until this study.

Range. Pliocene and Pleistocene, southern Australia (Western Australia, South Australia, New South Wales)⁶.

Wonambi barriei Scanlon, sp. nov.

Etymology. Named in honour of D. John Barrie, who has collected and prepared most of the material of *W. naracoortensis*.

Diagnosis. Trunk vertebrae less than 15 mm wide in adults (total length estimated to be less than 3 m); relatively large neural canal and shallow zygosphenoid; zygapophyses approximately 20° above horizontal. Single, large parazygantral foramen on each side of neural arch. Palatine relatively narrow (mediolateral dimension of choanal process subequal to that of tooth-bearing portion).

Material. Nine well-preserved vertebrae from WW Site, Riversleigh (northwest Queensland, Australia) represent most body regions, probably from a single individual (Queensland Museum F23038, holotype mid-trunk vertebra, and F40194, paratypes). Abundant postcranial material from various other Oligo-Miocene sites at Riversleigh is referred to this taxon, including a series of cloacal and anterior caudal vertebrae from CS Site (QM F40189). Referred cranial remains include associated jaw elements from CS Site (near-complete left palatine QM F40190 and fragments of the right palatine QM F40191, right and left pterygoids QM F23047 and 23048, articular and prearticular region of left mandible QM F23077, 23078), and fragments of two maxillae from CS (QM F39932) and WW Site (QM F40193).

Range. Late Oligocene to Early Miocene of Riversleigh, northwest Queensland¹³.

Comments. Elements of *W. barriei* are half the linear dimensions of corresponding elements of *W. naracoortensis*, implying a roughly eightfold difference in mass. Vertebrae of *W. barriei* can be interpreted as coming from adults of a small form because of the full extent of perichondral ossification, and the proportions of the centrum and neural arch. The two species are similar in these

proportions, whereas if the smaller forms merely represented juveniles of *W. naracoortensis* they would have shorter, broader and shallower centra, and dorsoventrally thinner neural arches¹⁵. *W. barriei* exhibits narrow palatine bones, an autapomorphy not found in *W. naracoortensis* or other madtsoiids¹⁴. It also lacks the autapomorphy of *W. naracoortensis*, small and numerous parazygantral foramina. Neither of these traits appears to be correlated with size in madtsoiids or other snakes. Each species spans a considerable stratigraphic interval without noticeable change in size or morphology.

The anatomy of *Wonambi* is revised here on the basis of newly identified elements (for example, frontal and ectopterygoid of *W. naracoortensis* and pterygoid of *W. barriei*) and a re-interpretation of known elements, which have only been briefly discussed and not yet studied phylogenetically⁵. No other madtsoiid taxon is represented by such complete material; therefore, *Wonambi* provides the best available evidence for the phylogenetic relationships of madtsoiids with other snakes. The known elements of other madtsoiids^{1,3,14} are similar to those of *Wonambi* and suggest that madtsoiids remained morphologically conservative throughout their long history.

The skull of *Wonambi* (Figs 1 and 2a–c) is long and low, the snout rounded, the palate very broad, and the braincase constricted behind the orbits but widening posteriorly. The parietal and frontal are most like those of *Dinilysia*⁸ in shape (including the large lateral crests of the parietal⁵) and attachments to each other and surrounding bones. The snout is rigidly attached to the frontal (Fig. 1) at a vertical frontonasal suture; the frontonasal (prokinetic) joint of modern snakes is thus poorly developed (condition unknown in *Pachyrhachis*⁷, in *Dinilysia* the prokinetic joint is described as having only limited mobility^{8,16}). The upper jaws have relatively tight and extensive connections to the central elements of the skull. The flat, pitted anterior end of the maxilla suggests a short ligamentous connection between the maxilla and premaxilla. The prefrontal is sutured to the maxilla but movably articulated with the frontal. There is an extensive palatine–vomer contact and small but distinct facets on the pterygoids for the basipterygoid articulations. A facet-like ‘trough’ on the suborbital region of the maxilla suggests retention of a jugal¹⁴ which was also present in *Pachyrhachis*⁷ and possibly *Dinilysia*⁸ but lost in all extant snakes. These primitive features indicate that kinesis of the upper jaws was more limited than in modern snakes, but similar to that of

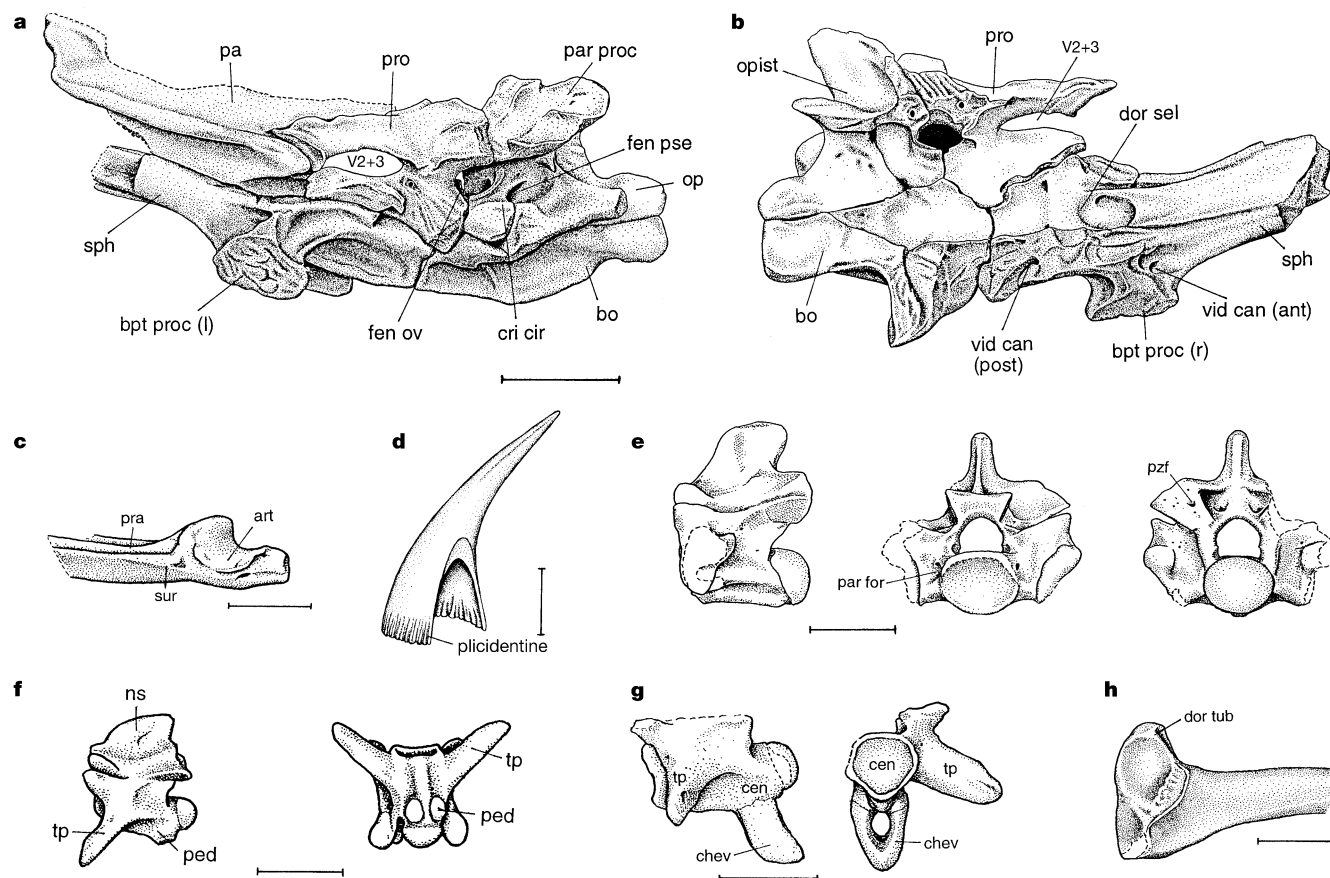


Figure 2 Selected elements of *Wonambi* exhibiting phylogenetically important characters. **a**, External view of braincase of *Wonambi naracoortensis* (SAMP30178) in left lateral view. Scale bar, 10 mm. **b**, Internal view of braincase in right lateral and slightly dorsal view. **c**, Articular region of right mandible of *W. barriei* (QMF23077) showing sutures between articular, surangular and prearticular. Scale bar, 3 mm. **d**, Reconstructed tooth, with part of base cut away, showing plicidentine on internal and external surfaces of tooth base (based on FU1762). Scale bar, 2 mm. **e**, Holotype mid-dorsal vertebra of *W. barriei* (QMF23038) in lateral, anterior and posterior views. Scale bar, 5 mm. **f**, Caudal vertebra of *W. barriei* (QMF40189) in lateral and ventral views, showing paired pedicels near posterior end of centrum. Scale bar, 4 mm. **g**, Partial caudal vertebra of *W. naracoortensis*

(SAMP30178) in lateral and anterior views showing articulated chevron. Scale bar, 5 mm. **h**, Head of left rib of *W. naracoortensis* (SAMP30178) in anterolateral view showing dorsal tubercle. Scale bar, 10 mm. Abbreviations as in Fig. 1 plus: art, articular; bpt process (l), left basipterygoid process; bpt process (r), right basipterygoid process; cen, centrum; chev, chevron; cri cir, crista circumfenestralis; dor sel, dorsum sellae; dor tub, dorsal tubercle; fen ov, fenestra ovalis; fen pseu, fenestra pseudorotunda; ns, neural spine; par for, paracotylar foramen; par proc, paroccipital process; ped, pedicel for chevron; pra, prearticular; pzf, parazygantral foramen; top, transverse process; vid can (ant), anterior opening of vidian canal; vid can (post), posterior opening of vidian canal; V2+3, trigeminal foramen.

anilioids¹⁷. However, the mandible was longer than the skull, and the supratemporals and/or quadrates must have projected posteriorly, as in *Pachyrhachis* and macrostomatan snakes, facilitating the swallowing of large items.

The anterior trunk region, representing about one third of all trunk vertebrae, comprises relatively high and narrow vertebrae with prominent single hypapophyses, and synapophyses projecting partly below the centrum. Mid-trunk vertebrae are broader, with slightly lower and longer neural spines, more dorsally placed synapophyses, and weak, posteriorly bifid or trifid haemal keels (Fig. 2e). More posteriorly the trunk vertebrae become smaller but relatively longer. Haemal keels disappear just before the cloacal region, but reappear on the cloacals and caudals. There are at least three cloacals with fused, forked lymphapophyses. At least one anterior caudal ('pygal') lacks haemapophyses, but other caudals have paired oval facets for their attachment on the posterior part of the centrum, near the ventral midline (Fig. 2f). The haemapophyses, which remain articulated to the centrum of one caudal vertebra in *W. naracoortensis*, are fused distally into a true chevron (Fig. 2g). Trunk ribs are curved throughout their length, with their

proximal facets slightly constricted between the strongly concave diapophyseal and flatter parapophyseal surfaces, and there is a prominent dorsal process (tuber costae) adjacent to the facet (Fig. 2h). No limb or girdle elements are yet known.

Madtsoiids are usually interpreted as an early appearing group of advanced snakes related to boas and pythons^{1,4,5,19}, although it has been suggested that they might represent a much more primitive lineage^{2,3}. Until recently, most fossil snake 'families' (including madtsoiids) were not represented by sufficient cranial material for rigorous assessment of relationships¹⁻³. The new information on *Wonambi* allows madtsoiids to be included in such an analysis for the first time. Accordingly, a phylogenetic analysis was performed including all living and fossil snake families (except those known from only postcranial elements), and incorporating the most extensive set of morphological characters so far assembled for snakes (see Methods). Very similar trees resulted whether multistate characters were ordered or unordered, and whether varanoid lizards, or amphisbaenians and dibamids, were used to infer character polarity. The results (Fig. 3) show that madtsoiids are among the most 'primitive' of snakes: as basal as the upper

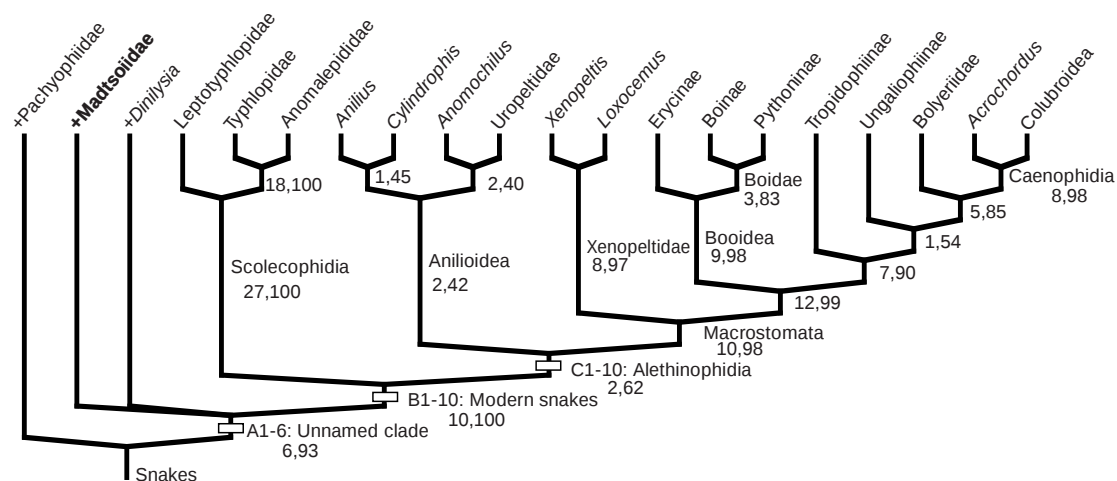


Figure 3 Cladogram (strict consensus of two most parsimonious trees, each with length = 649, consistency index = 0.49, retention index = 0.66) showing relationships among snake lineages and the very basal position of madtsoiids. This is based on a phylogenetic analysis of 234 morphological characters in 20 snake taxa (see Methods). Extinct taxa are denoted by '+'. The first number at each clade refers to Bremer support; the second refers to bootstrap frequency. Clade A, *Pachyrhachis* retains the following primitive characters shared with varanoid lizards (the derived state that unites madtsoiids, *Dinilysia* and modern snakes is listed afterwards in parentheses). A1, Exoccipitals not in contact above foramen magnum (Fig. 1a) (exoccipitals in contact, excluding supraoccipital from skull margin). A2, Parietal less than 40% of skull length, that is, premaxilla–occiput dimension (Fig. 1a) (more than 40%). A3, Sagittal dimension of supraoccipital long, more than 50% transverse dimension (Fig. 1a) (short, less than 50% transverse dimension). A4, Coronoid overlaps lateral surface of surangular (Fig. 2c) (does not overlap lateral surface). A5, Free forked cloacal ribs, that is, 'lymphapophyses', absent (present). A6, Rib heads without dorsal process (Fig. 2h) (process present). In addition, *Pachyrhachis* is more primitive than all living snakes in retaining a jugal bone, tibia, fibula and tarsals, in having a distinct cervical (neck) region with short narrow ribs, and in possessing fewer than 120 precloacal vertebrae^{7,18} (these primitive features might also characterize madtsoiids and/or *Dinilysia*, but the relevant regions in both are insufficiently known). Clade B, *Pachyrhachis* (where known), madtsoiids and *Dinilysia* retain the following primitive characters shared with varanoid lizards (the derived state that unites scolecophidian and alethinophidian snakes is listed afterwards in parentheses). B1, Ectopterygoid clasps dorsal and ventral surfaces of pterygoid (ectopterygoid overlaps ventral surface of pterygoid only). B2, Basipterygoid process prominent (Fig. 2) (basipterygoid processes reduced or lost; re-acquired in some booids). B3, Alar process of prootic long, narrow and distinct (Fig. 2) (alar process not distinct). B4, Crista circumfenestralis does not converge

to enclose stapedial footplate (Fig. 2a) (crista converges to partially or fully enclose stapedial footplate). B5, Paroccipital process very long (Fig. 2a) (reduced or absent). B6, Articular and surangular not fused in region of articular facet (Fig. 2c) (reduced or absent). B6, Articular and surangular not fused in region of articular facet (Fig. 2c) (articular and surangular fused). B7, Zygosphenal buttress with concave anterior edge between zygosphenes (Fig. 2e) (straight). B8, Paracotylar foramina present (Fig. 2e) (absent; reversals within Macrostromata). B9, Haemapophyses not fused to caudal centra (Fig. 2f, g) (fused to centra when present). B10, Haemapophyses are chevron-shaped, that is, joined distally (Fig. 2g) (separated distally when present). Madtsoiids are more primitive than *Dinilysia* and modern snakes in retaining plicidentine on the tooth bases (Fig. 2d), and lacking prezygapophysial processes (Fig. 2e); however, conflicting characters mean that the relationships between madtsoiids, *Dinilysia* and modern snakes remain unresolved. Clade C, *Pachyrhachis*, madtsoiids and *Dinilysia* share with scolecophidians the following primitive (varanoid-like) characters, which are modified in alethinophidians (derived state that unites alethinophidians is listed afterwards in parentheses). C1, Medial descending lamina of frontal absent (present). C2, Frontoparietal suture straight or M-shaped (Fig. 1a) (suture U-shaped and deeply concave anteriorly). C3, Vidian canal not opening intracranially (Fig. 2a, b) (opening intracranially). C4, Laterosphenoid absent (Fig. 2a, b) (laterosphenoid present). C5, Coronoid process formed entirely by coronoid (Fig. 1c) (coronoid process formed partly or entirely by surangular). In addition, scolecophidians are more primitive than alethinophidians in several soft anatomical traits which cannot be scored in the fossil taxa. C6, Undifferentiated ventral scales (midventral scales expanded transversely). C7, Presence of first branchial arch elements (absent). C8, *Adductor mandibulae externus pars superficialis* inserts on *adductor externus medialis* (inserts on *adductor externus profundus*). C9, Single pair of thymus glands (two pairs). C10, Non-lobed kidneys (lobed).

Cretaceous *Dinilysia*^{8,20} and only slightly more derived than the most basal known snake, the mid-Cretaceous *Pachyrhachis*⁷. In particular, all three fossil taxa lie outside the clade consisting of living snakes; this exclusion is supported by numerous compelling characters, a Bremer index of 10 and a bootstrap value of 100% (Fig. 3).

This analysis refutes the suggestion that *Pachyrhachis* is a derived rather than primitive snake²¹. Uniting *Pachyrhachis* with Macrostromata²¹ entails a significant increase in tree length (28 steps, $P < 0.0001$; see Methods) and implies reacquisition of hindlimbs and several skull bones. Other positions have also been suggested for madtsoiids (usually interpreted as booids^{1,4,5,19}) and *Dinilysia* (as an anilioid²⁰, booid¹ or basal alethinophidian^{12,22}). All these arrangements are significantly less parsimonious than the preferred tree (all entail between 10 and 26 extra steps and can be rejected at $P < 0.0001$). The arrangement obtained for living taxa is similar to those found in recent analyses^{1,12,22–24}. There is strong support for the monophyly of booids (boas, pythons and erylins), boids (boas and pythons) and xenopeltids (*Xenopeltis* and *Loxocemus*), clades which have been either poorly corroborated or contradicted in recent studies^{12,22–24}. There is also support for monophyly^{12,22} rather than paraphyly²³ of anilioids (*Anilius*, *Cylindrophis*, uropeltids and *Anomochilus*). This region of the tree is weak^{12,22,23}, however, as conflicting characters support nesting scoleophidians within anilioids, as the sister group of *Anomochilus*. Traits supporting this alternative arrangement, however, are almost entirely losses and reductions associated with fossoriality and miniaturization¹⁸. Within scoleophidians, a typhlopoid–anomalepidid clade^{12,23} is strongly supported over a typhlopoid–leptotyphlopoid clade²².

The small, worm-like scoleophidians are usually thought to be the most basal snakes^{1,11,12,22–24}, an arrangement consistent with the widely held idea that snakes evolved from tiny burrowing ancestors. This analysis indeed confirms that scoleophidians are the most basal living snakes; however, three fossil taxa emerge as even more basal than scoleophidians. Thus, the nearest relatives of snakes (mosasaurs and terrestrial varanoid lizards) and the three most basal snake lineages (*Pachyrhachis*, madtsoiids and *Dinilysia*) were all large predators with wide gapes. Mosasaurs and *Pachyrhachis* were marine, whereas madtsoiids and *Dinilysia* were either (surface-active) terrestrial or semi-aquatic^{1–6,14,16,25}, showing none of the extensive suite of traits correlated with fossoriality¹⁷. This challenges the widespread assumption that snakes arose through a small, burrowing ancestor^{9–12,23}: taxa straddling the lizard–snake transition are all large predators and either marine or (surface-active) terrestrial. These results suggest instead that the snake body form arose in a large, non-fossorial ancestor, perhaps for eel-like swimming²⁶ or sliding through dense vegetation^{27,28}. Fossoriality makes its debut later in snake evolution, in ‘modern’ forms such as scoleophidians and anilioids. □

Methods

Two-hundred and thirty-four morphological characters (191 osteological, 43 soft anatomical), adapted and expanded from recent analyses of snake phylogeny^{7,23,29}, were scored for all snake ‘families’. Characters were polarized by reference to the nearest lizard outgroups, mosasaurs and terrestrial varanoids¹⁸, and the tree rooted with a hypothetical ancestor coded with these inferred primitive states. Use of an amphibaenian–dibamid clade as the outgroup (see ref. 18) yielded similar trees. Character descriptions and data matrix in Nexus/PAUP format are available as Supplementary Information. Analyses were performed using PAUP version 4.0 (ref. 30). Bremer support was determined using converse constraints; bootstrap values are based on 1,000 heuristic replicates. Variability within terminal taxa was treated as uncertainty (regarding the primitive state) when calculating tree lengths. Two analyses were carried out: multistate characters ordered according to morphoclines where possible, as discussed in the character descriptions; and multistate characters all unordered. The branch-and-bound search found two most parsimonious cladograms in the ‘ordered’ analysis, with madtsoiids and *Dinilysia* forming either a clade or successive sister taxa to living snakes; the strict consensus of these is shown in Fig. 3. The ‘unordered’ analysis found a single most parsimonious cladogram corresponding to the first topology. Thus, the phylogenetic results are not dependent on assumptions of character state evolution (‘ordering’). Only unambiguous characters

(those invariant under accelerated and delayed optimization) are listed in the clade diagnoses in Fig. 3. To test alternative proposed positions for each fossil taxon, the ‘backbone constraints’ command was used: the fossil taxon in question was constrained to form a clade with its putative living relatives to the exclusion of other living taxa, but other fossil taxa were allowed to ‘float’. The most parsimonious tree(s) consistent with the backbone constraint was compared with the most parsimonious (unconstrained) trees using the Templeton and Kishino–Hasegawa tests in PAUP. Tree statistics, Bremer/ bootstrap support, and Templeton/Kishino–Hasegawa test values reported (Fig. 3) are those pertaining to the ‘ordered’ analysis; those in the ‘unordered’ analysis were almost identical.

Received 1 June; accepted 16 November 1999.

1. Rage, J. C. *Handbuch der Paläoherpologie. Teil 11. Serpentes* (Gustav Fischer, Stuttgart, 1984).
2. McDowell, S. B. in *Snakes: Ecology and Evolutionary Biology* (eds Seigel, R. A., Collins, J. T. C. & Novak, S. S.) 1–50 (Macmillan, New York, 1987).
3. Rage, J. C. Fossil snakes from the Paleocene of São José de Itaboraí, Brazil. Part I. Madtsoiidae, Aniliidae. *Palaeovertebrata* **27**, 109–144 (1998).
4. Smith, M. J. Small fossil vertebrates from Victoria Cave, Naracoorte, South Australia. IV. Reptiles. *Trans. R. Soc. S. Aust.* **100**, 39–51 (1976).
5. Barrie, D. J. Skull elements and associated remains of the Pleistocene boid snake *Wonambi naracoortensis*. *Mem. Qd Mus.* **28**, 139–151 (1990).
6. Scanlon, J. D. First records from Wellington Caves, New South Wales, of the extinct madtsoiid snake *Wonambi naracoortensis* Smith, 1976. *Proc. Linn. Soc. NSW* **115**, 233–238 (1995).
7. Lee, M. S. Y. & Caldwell, M. W. Anatomy and relationships of *Pachyrhachis problematicus*, a primitive snake with hindlimbs. *Phil. Trans. R. Soc. Lond. B* **353**, 1521–1552 (1998).
8. Estes, R., Frazzetta, T. H. & Williams, E. E. Studies on the fossil snake *Dinilysia patagonica* Woodward: Part 1. Cranial morphology. *Bull. Mus. Comp. Zool. Harv.* **140**, 25–74 (1970).
9. Walls, G. L. Ophthalmological implications for the early history of snakes. *Copeia* **1940**, 1–8 (1940).
10. Bellairs, A. d’A. & Underwood, G. The origin of snakes. *Biol. Rev.* **26**, 193–237 (1951).
11. Underwood, G. *A Contribution to the Classification of Snakes* (British Museum (Natural History), London, 1967).
12. Rieppel, O. A review of the origin of snakes. *Evol. Biol.* **22**, 37–130 (1988).
13. Archer, M., Hand, S. J., Godthelp, H. & Creaser, P. in *Actes du Congrès Biochron ’97* (eds Aguilar, J.-P., Legendre, S. & Michaux, J.) 131–152 (École Pratique des Hautes Études Institute de Montpellier, Montpellier, 1997).
14. Scanlon, J. D. *Nanowana* gen. nov., small madtsoiid snakes from the Miocene of Riversleigh: sympatric species with divergently specialised dentition. *Mem. Qd Mus.* **41**, 393–412 (1997).
15. LaDuke, T. C. The fossil snakes of Pit 91, Rancho La Brea, California. *Nat. Hist. Mus. LA County Contrib. Sci.* **424**, 1–28 (1991).
16. Frazzetta, T. H. Studies on the fossil snake *Dinilysia patagonica* Woodward. II. Jaw machinery in the earliest snakes. *Forma et Functio* **3**, 205–221 (1970).
17. Cundall, D. Feeding behaviour in *Cylindrophis* and its bearing on the evolution of alethinophidian snakes. *J. Zool.* **237**, 353–376 (1995).
18. Lee, M. S. Y. Convergent evolution and character correlation in burrowing reptiles: towards a resolution of squamate phylogeny. *Biol. J. Linn. Soc.* **65**, 369–453 (1998).
19. Underwood, G. in *Morphology and Biology of Reptiles* (eds Bellairs, A. d’A. & Cox, C. B.) 151–175 (Academic, London, 1976).
20. Woodward, A. S. On some extinct reptiles from Patagonia, of the genera *Miolania*, *Dinilysia*, and *Genyodectes*. *Proc. Zool. Soc. Lond.* **1901**, 169–184 (1901).
21. Zaher, H. The phylogenetic position of *Pachyrhachis* within snakes (Squamata, Lepidosauria). *J. Vert. Paleontol.* **18**, 1–3 (1998).
22. Kluge, A. G. Boine snake phylogeny and research cycles. *Misc. Publ. Mus. Zool. Univ. Michigan* **178**, 1–58 (1991).
23. Cundall, D., Wallach, V. & Rossman, D. S. The systematic relationships of the snake genus *Anomochilus*. *Zool. J. Linn. Soc.* **109**, 275–299 (1993).
24. Heise, P. J., Maxson, L. R., Dowling, H. G. & Hedges, S. B. Higher-level snake phylogeny inferred from mitochondrial DNA sequences of 12S rRNA and 16S rRNA genes. *Mol. Biol. Evol.* **12**, 259–265 (1995).
25. Hecht, M. K. The vertebral morphology of the Cretaceous snake, *Dinilysia patagonica* Woodward. *N. Jb. Geol. Paläont. Mh.* **1982**, 523–532 (1982).
26. Nopsca, F. *Eidosaurus* und *Pachyophis*. Zwei neue Neocom-Reptilien. *Palaeontographica* **65**, 99–154 (1923).
27. Cope, E. D. On the reptilian orders Pythonomorpha and Streptosauria. *Proc. Boston Soc. Nat. Hist.* **12**, 250–266 (1869).
28. Janensch, W. Über *Archaeophis proavus* Mass., eine Schlange aus dem Eocän des Monte Bolca. *Beitr. z. Paläont. Geol. Österreich-Ungarns* **19**, 1–33 (1906).
29. Wallach, V. & Rainer, G. Visceral anatomy of the Malaysian snake genus *Xenophidion*, including a cladistic analysis and allocation to a new family. *Amphibia-Reptilia* **19**, 385–404 (1998).
30. Swofford, D. L. *PAUP* Version 4—Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Computer program and documentation. (Sinauer, Sunderland, Massachusetts, 1999).

Supplementary information is available on Nature’s World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank D. J. Barrie, M. Archer, R. E. Molnar, N. Pledge and R. T. Wells for access to materials, and V. Wallach, G. Underwood, J.-C. Rage, D. J. Barrie, S. E. Evans, H. W. Greene, D. Cundall, M. W. Caldwell and A. G. Kluge for discussion. This research was supported by Australian Research Council grants to M.L. and J.S. Work at Riversleigh was supported by the Australian Research Council and University of New South Wales (to M. Archer), and work at Naracoorte by Flinders University, the South Australian Museum, L. and G. Henschke, the Barrie family and numerous volunteers.

Correspondence and requests for materials should be addressed to J.S. (e-mail: jscanlon@ultra.net.au).

Stimulation by ammonium-based fertilizers of methane oxidation in soil around rice roots

Paul L. E. Bodelier^{*,†}, Peter Roslev[‡], Thilo Henckel^{*} & Peter Frenzel^{*}

^{*}Max Planck Institute for Terrestrial Microbiology, Department of Biogeochemistry, Karl-von-Frischstrasse, D-35043 Marburg, Germany

[†]Centre for Limnology, Netherlands Institute of Ecology, Rijksstraatweg 6, 3631 AC, Nieuwersluis, The Netherlands

[‡]Environmental Engineering Laboratory, Aalborg University, Sohngaardsholmsvej 57, DK-9000, Aalborg, Denmark

Methane is involved in a number of chemical and physical processes in the Earth's atmosphere, including global warming¹. Atmospheric methane originates mainly from biogenic sources, such as rice paddies and natural wetlands; the former account for at least 30% of the global annual emission of methane to the atmosphere². As an increase of rice production by 60% is the most appropriate way to sustain the estimated increase of the human population during the next three decades³, intensified global fertilizer application will be necessary³; but it is known that an increase of the commonly used ammonium-based fertilizers can enhance methane emission from rice agriculture. Approximately 10–30% of the methane produced by methanogens in rice paddies is consumed by methane-oxidizing bacteria associated with the roots of rice^{4,5}; these bacteria are generally thought to be inhibited

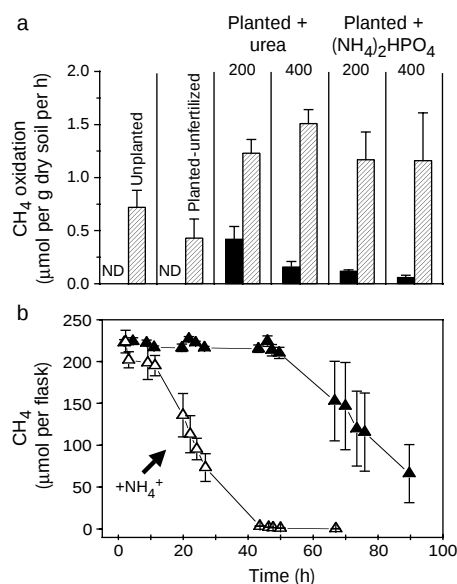


Figure 1 Methane-oxidizing activities in rice soil. Shown are the effects of ammonium-based fertilization and rice plants on methane oxidation rates of unplanted and rhizospheric soil from microcosms that were either unfertilized or supplemented with urea (200 or 400 kg N ha⁻¹) or (NH₄)₂HPO₄ (200 or 400 kg N ha⁻¹). The bars in **a** represent the arithmetic means of the initial (black bars) and induced (hatched bars) oxidation rates of four replicate microcosms ($\pm$ 1 s.d.). ND, assays without a detectable initial oxidation rate. **b**, The effect of ammonium addition on methane oxidation of rhizospheric soil slurries from unfertilized microcosms. Filled triangles, the non-supplemented slurries; open triangles, the methane depletion in the slurries supplemented with ammonium (end concentration, 2 mM). All values represent the arithmetic means of four replicate microcosms ($\pm$ 1 s.d.).

by ammonium-based fertilizers, as was demonstrated for soils^{6–8} and sediments^{9,10}. In contrast, we show here that the activity and growth of such bacteria in the root zone of rice plants are stimulated after fertilization. Using a combination of radioactive fingerprinting¹¹ and molecular biology¹² techniques, we identify the bacteria responsible for this effect. We expect that our results will make necessary a re-evaluation of the link between fertilizer use and methane emissions, with effects on global warming studies.

Methanotrophic bacteria utilize methane as the sole carbon and energy source; they are subdivided into types I and II, on the basis of phylogeny, physiology, morphology and biochemistry, including characteristic phospholipid ester-linked fatty acids (PLFA) in their cell membranes¹³. These bacteria cannot be active in flooded rice soils without oxygen derived from the rice plant. This is because of their obligatory aerobic metabolism. These obligatory aerobic bacteria live in close association with the roots of wetland plants (for example, rice) which provide them with necessary oxygen^{14,15} to oxidize methane which diffuses from the anoxic bulk soil to the rhizosphere. Here we examined the effect of fertilization on methane oxidation using 'microcosms' planted with rice; in each

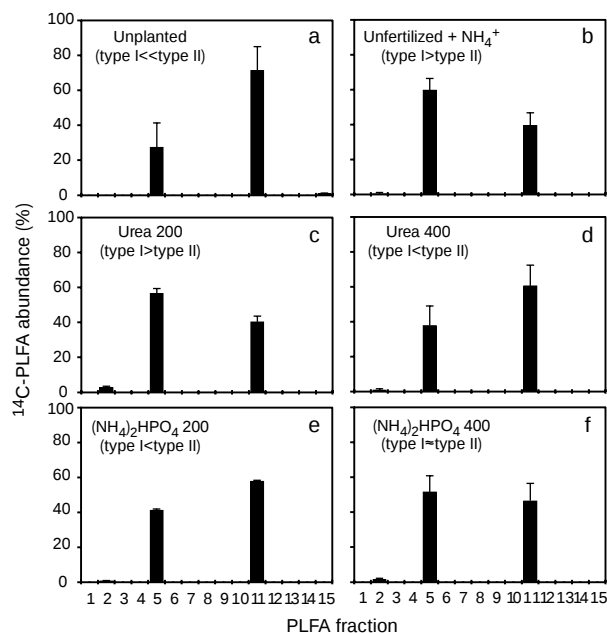
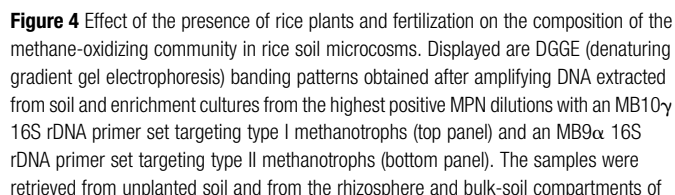
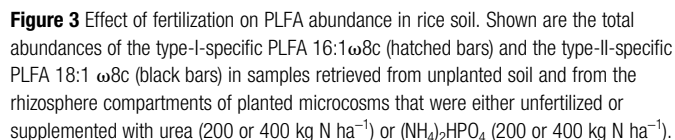


Figure 2 ¹⁴C-labelled phospholipid ester-linked fatty acid (¹⁴C-PLFA) fingerprints of methane-oxidizing bacteria from soil samples incubated with ¹⁴CH₄. The fingerprints were obtained from unplanted microcosms (**a**), and the rhizosphere compartments from planted microcosms that were supplemented with urea at 200 kg N ha⁻¹ (**c**) or 400 kg N ha⁻¹ (**d**), or with (NH₄)₂HPO₄ at 200 kg N ha⁻¹ (**e**) or 400 kg N ha⁻¹ (**f**). **b**, The fingerprint from soil originating from the rhizosphere compartment of unfertilized microcosms that were supplemented with ammonium (compare Fig. 1b) in post-harvest slurry incubations. The graphs show the percentage of radioactivity detected in each PLFA fraction. ¹⁴CH₄ incorporation into PLFAs that eluted mainly in fraction 5 (for example, 16:1ω8, 16:1ω7, 16:1ω5 and 16:0) indicate methane oxidation by type I methane oxidizers, whereas radioactive PLFAs detected mainly in fraction 11 (for example, 18:1ω9, 18:1ω8, 18:1ω7, and 18:0) indicate oxidation dominated by type II methane oxidizers. Bars indicate the mean ($\pm$ 1 s.d.) of two subsamples taken from 1 out of 4 replicate microcosms per treatment. The fingerprint in **b** corresponds to the ammonium-supplemented slurry shown in Fig. 1. The metabolism of ¹⁴CH₄ was defined as dominated by type I methane-oxidizing bacteria when the ratio of fraction 5: fraction 11 was greater than 1.1. If this ratio was lower than 1.1, type II methanotrophs were assumed to dominate. This differentiation is based on laboratory experiments with pure cultures of type I and II methane-oxidizing bacteria²⁷.

fertilization (Kruskall–Wallis test, $P<0.001$) and the presence of rice plants (Kruskall–Wallis test, $P<0.05$) (Fig. 1a). With the unfertilized samples it took 50 h before any methane depletion could be detected (Fig. 1b, filled triangles). However, ammonium addition to parallel soil slurries from unfertilized microcosms led to an immediate activation of the methane-oxidizing bacteria (Fig. 1b, open triangles), demonstrating the essential role of ammonium in the consumption of methane in rice soil.

The bacteria responsible for the observations of accelerated methane oxidation were identified using radioactive fingerprinting and molecular biology techniques. In all treatments, radiolabelled methane was incorporated into PLFA fractions that contained fatty acids typical of type I and II methane-oxidizing bacteria (Fig. 2). Type II methane oxidizers dominated methane metabolism in unplanted, unfertilized soil, as indicated by recovery of ^{14}C -PLFAs mainly in fraction 11 (Fig. 2a). In contrast, long term-fertilization of planted soil with either urea or $(\text{NH}_4)_2\text{HPO}_4$ resulted in the activity of both type I and II methane oxidizers in the rhizosphere (Fig. 2b–f). The apparent activation of type I methane-oxidizing bacteria by fertilization was confirmed by the ^{14}C -PLFA fingerprints obtained after addition of ammonium to soil slurries retrieved from the rhizosphere (Fig. 2b). More radioactivity was recovered in PLFA



422

fraction 5 after addition of ammonium to such samples from unfertilized microcosms (Fig. 2b).

The total abundance of the PLFAs specific for methanotrophic bacteria (16:1 ω 8c for type I and 18:1 ω 8c for type II) can be regarded as a measure of methanotrophic biomass¹⁶. (Fatty acids are designated by the total number of carbon atoms. The degree of unsaturation is indicated by a number separated from the chain length by a colon and is followed by ω x, where x indicates the position of the double bond nearest to the aliphatic end (ω). The c indicates a *cis*-position of the double bond on the molecule.) Our data revealed that fertilizer addition to planted microcosms resulted in a 2–3-fold increase in the abundance of PLFAs specific for methanotrophic bacteria (Fig. 3). This was supported by using a most probable number approach where higher (Kruskal–Wallis test, $P < 0.001$) numbers of culturable methanotrophs were found in fertilized microcosms¹⁷. Fertilization resulted mainly in growth of type I methane oxidizers, as indicated by a ninefold increase in the type-I-specific PLFA biomarker in microcosms fertilized with 400 kg N ha⁻¹ of urea or (NH₄)₂HPO₄ (Fig. 3). The type-II-specific PLFA 18:1 ω 8c increased only 2–3-fold after fertilization, but this PLFA remained the dominant methanotrophic biomarker in fertilized samples. The dominance was most pronounced in unplanted and unfertilized samples where the type-II-specific biomarker was 7 times more abundant than type I biomarkers (Fig. 3).

By combining the PLFA profiles and denaturing gradient gel electrophoresis (DGGE) separation of 16S rDNA amplified by the polymerase chain reaction (PCR), a higher phylogenetic resolution

of the methane-oxidizing community in rice soil was obtained. The DGGE banding patterns (Fig. 4) and the respective sequences (Fig. 5) demonstrate that type and dose of fertilizer had little effect on the species composition of the methane-oxidizing community. Type I bands (RRI 1–6, RBI 1) were most closely related to *Methylobacter* species (Fig. 5), whereas type II bands (RRII 3–7) were related to *Methylosinus* and *Methylocystis* species. All the main bands, methanotrophic as well as non-methanotrophic, were highly related to sequences and clones from rice-field soil incubations¹² (Fig. 5). DGGE bands related to type II methane oxidizers were equally intense in all samples, whereas type-I-related DGGE bands were distinctly more intense in the rhizosphere samples (Fig. 4). These intensity differences are the consequence of differences in the original number of target sequences, and hence of the numbers of the respective methane oxidizers in the soil¹². Thus, as well as ammonium, the presence of the rice plant is an essential factor for type I methane-oxidizing bacteria to proliferate. This finding is supported by the PLFA data.

Enrichments from the highest positive dilutions of most probable number (MPN) counts contained only type II methanotrophs (Fig. 4, lower panel). These bands (ERRII 1–3, ERBII 1–3) matched the dominant bands obtained from soil samples, confirming the numerical dominance of type II methanotrophs in rice soil that was indicated by the PLFA analysis (Fig. 3).

Our results show that ammonium-based fertilization does not necessarily inhibit the consumption of methane in soils and sediments as has been repeatedly reported^{16–10}. The absence of inhibition

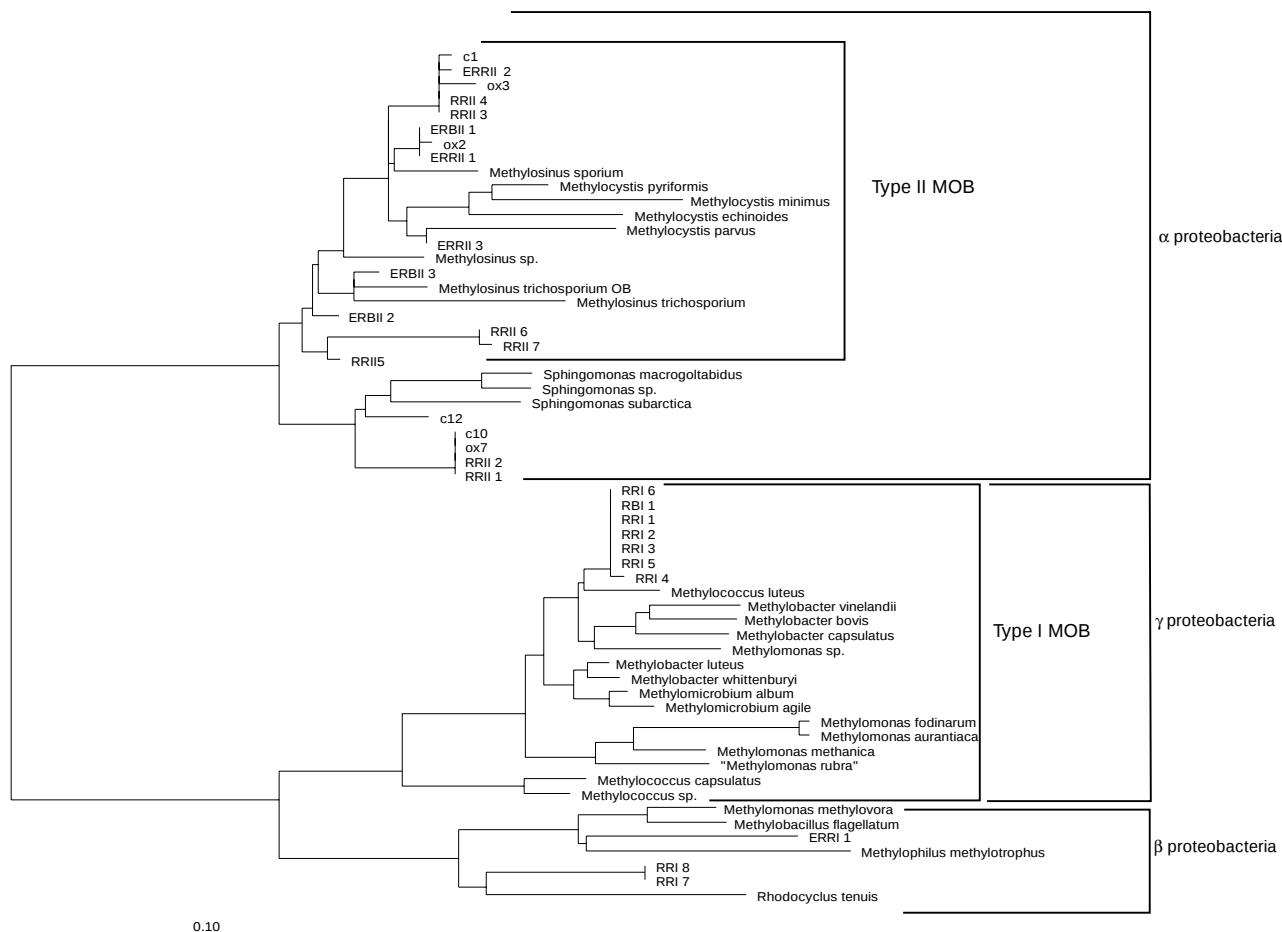


Figure 5 Phylogenetic tree. Shown is the relationship of partial 16S rDNA sequences retrieved from DGGE gels (see Fig. 4) with the most closely related members of the α -, β - and γ -proteobacteria. The scale bar indicates the estimated number of base changes per

nucleotide sequence position. 'c' and 'ox' are clones from rice field soil incubated without and with methane, respectively¹². MOB, methane oxidizing bacteria.

may be explained by the high methane availability¹⁷ in rice paddies, which counterbalances possible competitive inhibition by ammonium¹⁸. Furthermore, fertilizer-associated nitrite toxicity⁷ or salt effects^{19,20} on methane oxidation may be eliminated by plant uptake of these ions. Such a sink for toxic substances is missing in upland soils where methane concentrations are low, explaining the reported inhibition of methane oxidation by ammonium in soils and soil incubations^{6–10}.

Both type I and II conventional methanotrophs need ammonium to be active and to grow in the rice rhizosphere. However, the number of type I methane oxidizers was increased to a greater extent relative to the number of type II methane-oxidizing bacteria upon fertilization. Nitrogen availability has been suggested to select for type I methane oxidizers¹³, but experimental evidence has been missing until now.

Our data suggest that fertilizer application may lead to a reduction, rather than to an increase, of methane emission from wetland rice fields. Methane emission from the microcosms used in this study was reduced up to 57% by fertilizer application¹⁷. Moreover, a significant negative correlation has been found between porewater ammonium concentrations and methane emission, indicating increased methane oxidation to be the reason for this reduction¹⁷. However, we¹⁷ and other investigators^{21,22} have shown that fertilization can also enhance methane production due to higher plant biomass and subsequent carbon availability for methanogens. The net emission will be the balance between enhanced consumption and production. We expect that the results we report here will therefore lead to a re-evaluation of fertilizer practice and studies on methane emission from soils and sediments. □

Methods

Microcosms

Experimental system and design have been described^{17,23}. The microcosms had a physical separation between rhizosphere and bulk soil, and were incubated in growth cabinets for a period of 90 d at 70% relative humidity and in a light/dark cycle of 12/12 h at a photosynthetically active radiation of 450 $\mu\text{E m}^{-2} \text{s}^{-1}$ (99 W m^{-2}) and a temperature regime of night/day of 20 °C/25 °C. The surface of the soil was always covered with 2 cm of demineralized water. Soil and seeds (*Oryza sativa* var. Roma, type japonica) originated from an Italian rice paddy (Vercelli, northern Italy). The microcosms were either unplanted, planted and unfertilized, or planted and supplemented with urea (200 or 400 kg N ha^{-1}) or $(\text{NH}_4)_2\text{HPO}_4$ (200 or 400 kg N ha^{-1}). The fertilizer was added twice per week by injection into the soil with a syringe. After 90 d of incubation, the soil from the rhizosphere and bulk soil compartments was collected and soil slurries were prepared²⁴ for performing the analyses described below.

Methane oxidation rates

These (Fig. 1a) were determined by transferring subsamples of the above slurries to flasks that were supplemented with CH_4 (10,000 p.p.m.v.). CH_4 depletion was monitored by gas chromatographic (GC) analysis of the headspace gas as described earlier²⁴. From the sigmoidal CH_4 depletion curves (Fig. 1b, open triangles) initial (0–12 h) and induced (12–24 h) oxidation rates were calculated by linear regression. To demonstrate the nitrogen-limitation of methane oxidation in unfertilized microcosms, slurries from unfertilized microcosms were supplemented with $(\text{NH}_4)_2\text{SO}_4$ (final concentration 2 mM N).

¹⁴C-PLFA profiles

Directly after collecting the soil and preparation of soil slurries, 20 ml of slurry was transferred to a polyethersulphon membrane (pore size 0.2 μm , diameter 4.7 mm) by vacuum filtration. The membranes (with ~4 g of fresh soil) were transferred to 150-ml flasks, supplemented with 0.15 MBq ¹⁴ CH_4 and incubated at 25 °C for 3 d. PLFAs were extracted as described elsewhere¹¹. Radiolabelled PLFAs were separated into 15 fractions by capillary GC and collected as ¹⁴ CO_2 after combustion in the flame ionization detector. The radioactivity in each fraction was determined by liquid scintillation counting.

PCR and DGGE of soil DNA

DNA was extracted from soil and amplified by PCR as described¹². The primer sets MB10 γ and MB9 α amplify 16S rDNA of methylobacterial bacteria belonging to the γ - and α -proteobacteria¹², respectively. PCR products were separated by DGGE using a 35–70% and 45–70% denaturant gradient for MB10 γ and MB9 α PCR products, respectively. For pure cultures and enrichments, 10 μl of PCR product was loaded onto the gel, whereas for soil samples 45 μl was applied. Bands were excised from the gels, reamplified and sequenced as described¹². The phylogenetic position of the derived sequences are shown in Fig. 4.

Phylogeny

16S rDNA sequences were aligned and placed phylogenetically with the ARB software package²⁵. Evolutionary distances between pairs of sequences were calculated by using the Jukes–Cantor and Felsenstein algorithms in the ARB software. The phylogenetic trees were constructed using the neighbour-joining algorithm supplied with the ARB package, based on complete 16S rDNA sequences of α -, β - and γ -proteobacteria. The partial 16S rDNA sequences retrieved from DGGE gels were added to the tree by keeping its topology constant²⁶. Sequences of partial 16S rRNA gene fragments of excised DGGE bands have been deposited in the GenBank database under accession nos AF179599–AF179610.

Received 31 August; accepted 9 November 1999.

- Crutzen, P. J. in *Ruminant Physiology: Digestion, Metabolism, Growth and Reproduction* (eds von Engelhardt, W., Leonhardt-Marek, S., Breves, G. & Giesecke, D.) 291–315 (Enke, Stuttgart, Germany).
- Neue, H. U. Fluxes of methane from rice fields and potential for mitigation. *Soil Use Mgmt* **13**, 258–267 (1997).
- Cassman, K. G. *et al.* Opportunities for increased nitrogen-use efficiency from improved resource management in irrigated rice systems. *Field Crops Res.* **56**, 7–39 (1998).
- Denier van der Gon, H. A. C. & Neue, H. U. Oxidation of methane in the rhizosphere of rice plants. *Biol. Fertil. Soils* **22**, 359–366 (1996).
- Bosse, U. & Frenzel, P. Activity and distribution of methane-oxidizing bacteria in flooded rice soil microcosms and in rice plants (*Oryza sativa*). *Appl. Environ. Microbiol.* **63**, 1199–1207 (1997).
- Stedler, P. A., Bowden, R. D., Mellilo, J. M. & Aber, J. D. Influence of nitrogen fertilization on methane uptake in temperate forest soil. *Nature* **341**, 314–316 (1989).
- King, G. M. & Schnell, S. Effect of increasing atmospheric methane concentration on ammonium inhibition of soil methane consumption. *Nature* **370**, 282–284 (1994).
- Gulledge, J., Doyle, A. P. & Schimel, J. P. Different NH_4^+ -inhibition patterns of soil CH_4 -oxidizer populations across sites. *Soil Biol. Biochem.* **29**, 13–21 (1997).
- Bosse, U., Frenzel, P. & Conrad, R. Inhibition of methane oxidation by ammonium in the surface layer of a littoral sediment. *FEMS Microbiol. Ecol.* **13**, 123–134 (1993).
- Van der Nat, F. J. W. A., DeBrouwer, J. F. C., Middelburg, J. J. & Laanbroek, H. J. Spatial distribution and inhibition by ammonium of methane oxidation in intertidal freshwater marshes. *Appl. Environ. Microbiol.* **63**, 4734–4740 (1997).
- Roslev, P., Iversen, N. & Henriksen, K. Direct fingerprinting of metabolically active bacteria in environmental samples by substrate specific radiolabelling and lipid analysis. *J. Microbiol. Methods* **31**, 99–111 (1998).
- Henckel, T., Friedrich, M. & Conrad, R. Molecular analysis of the methane-oxidizing microbial community in rice field soil by targeting the genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. *Appl. Environ. Microbiol.* **65**, 1980–1990 (1999).
- Hanson, R. S. & Hanson, T. E. Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439–471 (1996).
- King, G. M. Associations of methanotrophs with the roots and rhizomes of aquatic vegetation. *Appl. Environ. Microbiol.* **60**, 3220–3227 (1994).
- Gilbert, B. & Frenzel, P. Rice roots and CH_4 oxidation: the activity of bacteria, their distribution and the microenvironment. *Soil Biol. Biochem.* **30**, 1903–1916.
- Sundh, I., Borgå, P., Nilsson, M. & Svensson, B. H. Estimation of cell numbers of methanotrophic bacteria in boreal peatlands based on analysis of specific phospholipid fatty acids. *FEMS Microbiol. Ecol.* **18**, 103–112 (1995).
- Bodelier, P. L. E., Hahn, A. P., Arth, I. R. & Frenzel, P. Effects of ammonium-based fertilisation on microbial processes involved in methane emission from soils planted with rice. *Biogeochemistry* (submitted).
- Dunfield, P. F. & Knowles, R. Kinetics of methane oxidation by nitrate, nitrite, and ammonium in a humisol. *Appl. Environ. Microbiol.* **61**, 3129–3135 (1995).
- King, G. M. & Schnell, S. Effects of ammonium and non-ammonium salt additions on methane oxidation by *Methylosinus trichosporium* OB3b and maine forest soil. *Appl. Environ. Microbiol.* **64**, 253–257 (1998).
- Gulledge, J. & Schimel, J. P. Low-concentration kinetics of atmospheric CH_4 oxidation in soil and mechanism of NH_4^+ inhibition. *Appl. Environ. Microbiol.* **64**, 4291–4298 (1998).
- Lindau, C. W., Bollich, P. K., Delaune R. D., Patrick, W. H. Jr & Law, V. J. Effects of urea fertilizer and environmental factors on CH_4 emissions from a Louisiana, USA rice field. *Plant Soil* **136**, 195–203 (1991).
- Banik, A., Sen, M. & Sen, S. P. Effects of inorganic fertilizers and micronutrients on methane production from wetland rice (*Oryza sativa* L.). *Biol. Fertil. Soil* **21**, 319–322 (1996).
- Bodelier, P. L. E., Wijnhuizen, A. G., Blom, C. W. P. M. & Laanbroek, H. J. Effects of photoperiod on growth and denitrification by *Pseudomonas chlororaphis* in the root zone of *Glyceria maxima*, studied in a gnotobiotic microcosm. *Plant Soil* **190**, 91–103 (1997).
- Bodelier, P. L. E. & Frenzel, P. Contribution of methanotrophic and nitrifying bacteria to CH_4 and NH_4^+ oxidation in the rhizosphere of rice plants as determined by new methods of discrimination. *Appl. Environ. Microbiol.* **65**, 1826–1833 (1999).
- Strunk, O. & Ludwig, W. ARB: A Software Environment for Sequence Data (Technische Universität München, Munich, Germany, 1996); available at <http://www.biol.chemie.tu-muenchen.de/pub/ARB>.
- Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568 (1998).
- Roslev, P. & Iversen, N. Radioactive fingerprinting of microorganisms that oxidize atmospheric methane in different soils. *Appl. Environ. Microbiol.* **65**, 4064–4070 (1999).

Acknowledgements

We thank R. Conrad and P. Dunfield for comments on the manuscript, M. Friedrich for supporting the molecular studies, and B. Wagner and S. Fleisner for technical assistance. The project was supported by the EU, the Danish Technical Research Council and the Deutsche Forschungsgemeinschaft.

Correspondence and request for materials should be addressed to P.L.E.B. at the Netherlands Institute of Ecology (e-mail: bodelier@c1.nioo.knaw.nl).

Activin- and Nodal-related factors control antero-posterior patterning of the zebrafish embryo

Bernard Thisse*, Christopher V. E. Wright† & Christine Thisse*

* Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, CU de Strasbourg, France

† Department of Cell Biology, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA

Definition of cell fates along the dorso-ventral axis depends on an antagonistic relationship between ventralizing transforming growth factor- β superfamily members, the bone morphogenetic proteins¹ and factors secreted from the dorsal organizer, such as Noggin and Chordin. The extracellular binding of the last group to the bone morphogenetic proteins prevents them from activating their receptors^{2–4}, and the relative ventralizer:antagonist ratio is thought to specify different dorso-ventral cell fates. Here, by taking advantage of a non-genetic interference method using a specific competitive inhibitor, the Lefty-related gene product Antivin⁵, we provide evidence that cell fate along the antero-posterior axis of the zebrafish embryo is controlled by the morphogenetic activity of another transforming growth factor- β superfamily subgroup—the Activin and Nodal-related factors^{6–9}. Increasing *antivin* doses progressively deleted posterior fates within the ectoderm, eventually resulting in the removal of all fates except forebrain and eyes. In contrast, overexpression of *activin* or *nodal*-related factors converted ectoderm that was fated to be forebrain into more posterior ectodermal or mesodermal fates. We propose that modulation of intercellular signalling by Antivin/Activin and Nodal-related factors provides a mechanism for the graded establishment of cell fates along the antero-posterior axis of the zebrafish embryo.

According to the fate map of the zebrafish late blastula^{10,11} and specific gene expression patterns⁵, antero-posterior (A-P) patterning appears to be already established during late blastula/early gastrulation. At this stage, the embryo consists of an array of presumptive territories, from the margin to the animal pole: endoderm, mesoderm, spinal cord, hindbrain, midbrain and forebrain (Fig. 1). After gastrulation, the linear A-P organization of the ectoderm is essentially collinear with the arrangement of fates along the animal-vegetal axis seen at blastula stages (Fig. 1m). In contrast, the A-P organization of mesendoderm in post-gastrula embryos differs greatly from the blastula stage: the most vegetally cells that invaginate first during gastrulation form the leading edge of the mesendoderm and will therefore be the most anterior cells within this germ layer. Then, because of epiboly and convergence movements, structures derived from the ventral margin become located more posteriorly than those from the dorsal margin.

Activin has been implicated as the mesoderm-inducing signal, consistent with its ability to induce a variety of mesoderm derivatives in a dose-dependent manner^{6,12}. The activity of the Activin signalling pathway can be competitively inhibited by Antivin (Atv), a Lefty/EBAF (endometrial bleeding associated factor)-related factor⁵, which is expressed at the right time and place to limit mesendoderm induction at the blastula stage (Fig. 1a–c). Hence, we used Atv properties to carry out loss-of-function experiments to address directly the requirement in patterning of the A-P axis for inducers in the transforming growth factor- β (TGF β) superfamily, including Activin and Nodal-related factors. We injected *atv* RNA at the 1–4-cell stage to deliver the inhibitor into all cells. Low amounts of *atv* RNA (5 pg) caused the depletion of the most vegetally located cell fates, including the endoderm, the cephalic mesoderm and the

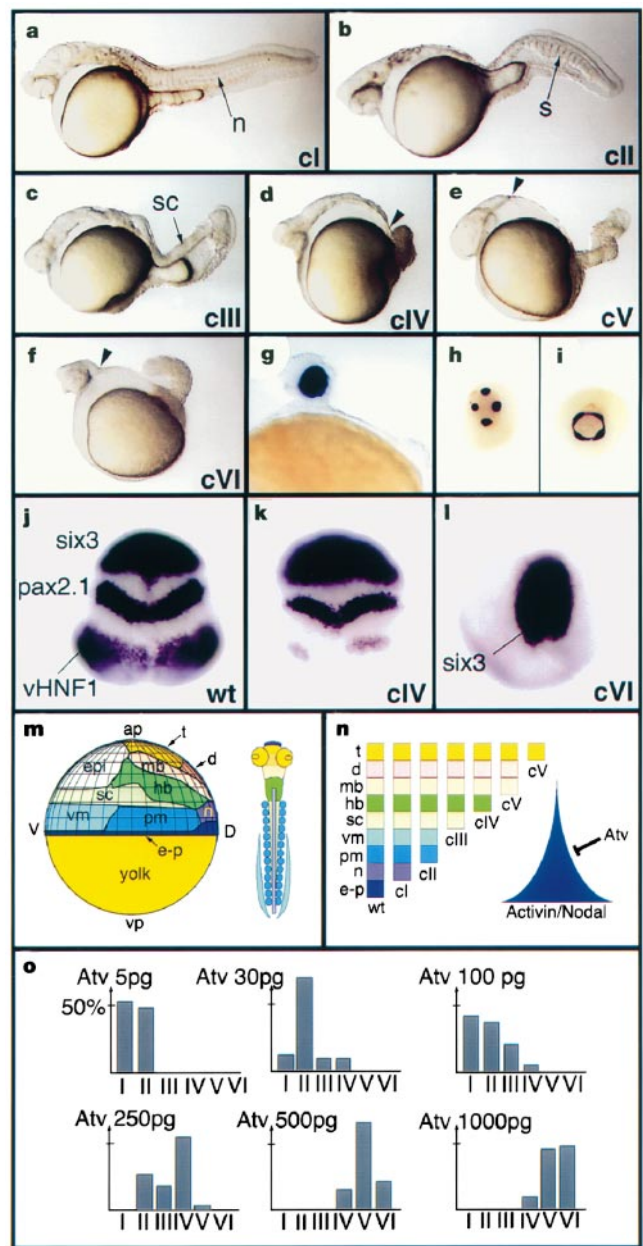


Figure 1 Dose-dependent depletion of structures along the animal-vegetal axis mediated by Antivin (Atv). Morphological defects induced by *atv* overexpression at 30 h (**a–i**), and analysis at gastrulation stage (90% epiboly) with a combination of ectodermal markers (**j–l**). **a**, cl phenotype lacking endoderm and prechordal plate. **b**, cII phenotype lacking additionally notochord (n). **c**, cIII phenotype devoid of mesoderm. **d**, cIV phenotype additionally lacking spinal cord (sc). **e**, cV phenotype lacking mesoderm, spinal cord and hindbrain. **f**, cVI embryos display only forebrain and one cyclopic eye. **g**, 35% of cVI embryos have a single lens, whereas 65% have multiple lens (**h**) or a circular lens placode marked by a lens-specific probe (SH372) (**i**). **j**, Wild-type (wt) gastrula labelled with *six3* (forebrain), *pax2.1* (midbrain–hindbrain boundary) and *vHNF1* (posterior hindbrain and anterior spinal cord). **k**, cIV phenotype lacking the spinal cord expression of *vHNF1*. **l**, cVI phenotype displaying only the *six3* expression domain. **m**, Left, drawing of the fate map of a zebrafish late blastula; right, derived structures at segmentation stages. Marginal cells give rise to mesendodermal fates: endoderm and prechordal plate (e–p), notochord (n), paraxial mesoderm (pm), ventral and lateral mesoderm (vm). In the ectoderm, from the margin to the animal pole, cells give rise to spinal cord (sc), hindbrain (hb), midbrain (mb), forebrain comprising diencephalon (d) and telencephalon (t). epi, epidermis; V, ventral; D, dorsal; vp, vegetal pole; ap, animal pole. **n**, In reference to the fate map, representation of the different phenotypes obtained after *atv* overexpression, from cl to cVI. **o**, Frequency of the different phenotypic classes ranked by *atv* RNA dose. s, somites. Arrowheads in **d–f** show the extent of the depletion mediated by Atv.

prechordal plate (cI phenotype; Fig. 1a). Increased *atv* doses (30 pg and 100 pg) caused the additional depletion of axial mesoderm, paraxial and ventral mesoderm (cII, cIII; Fig. 1b, c). Mesodermal territories were also deleted in a dose-dependent manner progressively from anterior to posterior. For example, in the dorsal region, the prechordal plate (expressing *gooseoid*¹³) which invaginates before notochordal cells (expressing *flh*¹⁴) was lost first. Similarly, anterior notochord/anterior somites were deleted using a lower *atv* RNA dose than caudal notochord/caudal somites (data not shown). In the absence of these marginal-zone fates, the embryo develops as an enlarged animal territory differentiating into structures fated to become ectoderm, but displaying an unexpectedly large degree of correct A–P organization (cIII; Fig. 1c).

We then tested whether even stronger *Atv*-mediated antagonization would also affect the patterning of the ectodermal layer. Embryos injected with 250 pg *atv* RNA, in addition to lacking mesoderm, failed to form spinal cord (cIV; Fig. 1d) and displayed only brain and an undifferentiated tail. Injection of 500 pg *atv* RNA led to the development of embryos additionally lacking hindbrain and ears (cV; Fig. 1e). At the highest dose of *atv* RNA injected (1,000 pg), embryos differentiated a single eye and part of the forebrain (cVI; Fig. 1f, g); higher doses did not enhance this phenotype. We also noted that, in embryos showing maximal *Atv*-mediated removal of posterior ectodermal fates, the cyclopic eye often contained multiple lenses (Fig. 1h) or formed a circular lens placode (Fig. 1i), indicating a complete loss of A–P polarity. This progressive depletion of ectodermal territories was already detectable at gastrulation, as shown by a combination of molecular markers (Fig. 1j–l).

In summary, inhibition of Activin-like signalling through *Atv* not only abolishes mesoderm induction⁵, but also results in a gradual posterior-to-anterior dose-dependent depletion of ectodermal subpopulations. The insensibility of structures derived from the animal pole to the inhibitory effects of *Atv* indicates that anterior ectoderm is the default fate under these circumstances.

Two TGF β superfamily members other than Activin that can dose-dependently induce different subpopulations of mesoderm were recently identified in zebrafish: *Znr1/Cyclops* and

Table 1 Activin and Nodal-related factors rescue posterior structures deleted by overexpression of *Activin*

	Phenotype					
	n	cVI (%)	cV (%)	cIV (%)	cIII (%)	cII (%)
<i>atv</i> (1 ng)	134	51	45	4	0	0
+ <i>activin</i> β B (0.05 pg)	112	12	76	8	4	0
+ <i>activin</i> β B (0.12 pg)	126	0	0	18	21	59
+ <i>activin</i> β B (0.25 pg)	116	0	0	23	25	47
+ <i>activin</i> β B (0.50 pg)	114	0	0	0	0	56
+ <i>znr2</i> (12.5 pg)	146	34	62	2	2	0
+ <i>znr2</i> (25 pg)	132	3	14	27	19	37
+ <i>znr2</i> (50 pg)	107	2	5	18	27	48
+ <i>znr2</i> (125 pg)	151	1	6	20	23	45
+ <i>znr1</i> (12.5 pg)	109	42	57	1	0	0
+ <i>znr1</i> (25 pg)	140	3	38	25	19	15
+ <i>znr1</i> (50 pg)	129	4	4	23	42	27
+ <i>znr1</i> (125 pg)	111	2	9	18	37	31

The resulting phenotypes of embryos injected with 1 ng *atv* and the indicated amount of *activin* or *NRs* RNA were scored after *in situ* hybridization with a combination of probes specific for structures formed at various positions along the A–P axis (see Fig. 2).

Znr2/Squint^{7–9}. Spatiotemporal analysis of the expression of these genes has revealed a large amount of overlap with that of *atv*. Moreover, *cyc/sqt* double mutant embryos⁹ as well as maternal-zygotic *oep* mutants thought to be involved in the *cyc/sqt* signalling pathway¹⁵ display a failure of mesoderm formation similar to that seen in cII phenotype *atv*-injected embryos (Fig. 1b). This suggests that the Nodal-related factors (NRs) are potential endogenous targets for *Atv*. In co-injection experiments (Fig. 2, Table 1), we showed that increasing the amounts of *Activin*/NRs dose-dependently and progressively rescued the A–P patterning defects from class cVI back to class cI phenotype. We conclude that *Atv*, as well as being an *Activin* antagonist, acts also as an anti-Nodal, consistent with the characterization of *Atv*/Lefty activities in other vertebrate species¹⁶.

In summary, the most vegetal fates of the blastula, the endoderm and prechordal plate are the first to be depleted by *Atv* action, whereas regions closest to the animal pole are the last to be affected and the first to be rescued by co-injection of *Activin*/NRs. These observations indicate that A–P patterning may be established at the blastula stage by a morphogenetic gradient of *Activin*/NRs; maximal activity occurs at the margin, and activity progressively decreases to the animal pole, with each territory in between being determined by the level of stimulation by the *Activin*/Nodal signalling pathway (Fig. 1n).

If the animal pole cell fates are associated with the absence of *Activin*/Nodal signalling, are the cells in this territory still capable of being transformed to other A–P fates by the local expression of *Activin*/NRs? To address this question, we targeted cells of the presumptive animal pole at the 128-cell stage with various concentrations of *Activin*/NRs (Fig. 3a). The clone of cells derived from the injected blastomere was followed using green fluorescent protein (GFP) as a lineage tracer (Fig. 3c, e). The injected cells appeared to be located in the cephalic region and were associated with various ectopic structures. As shown morphologically (Fig. 3d) or with different markers, the induced structures were of mesodermal or ectodermal identity. Injection of 30 fg *znr1* RNA at 24 h led to the formation of an ectopic notochord in place of the forebrain (Fig. 3f, g), induction of which was already detectable at gastrulation (Fig. 3h–k), and 8 fg replaced forebrain or eye by hindbrain (Fig. 3l–p). Similar results were obtained when targeting the animal pole with *activin* or *znr2* RNA (data not shown). Thus, the local expression of *Activin*/NRs can lead to the dose-dependent re-specification of animal pole derivatives (forebrain) into more vegetally derived structures. These results are consistent with studies in which the transplantation of lineage-labelled germring cells to the animal pole non-autonomously changed the presumptive forebrain into a patterned hindbrain¹⁷. On the basis of our results, and

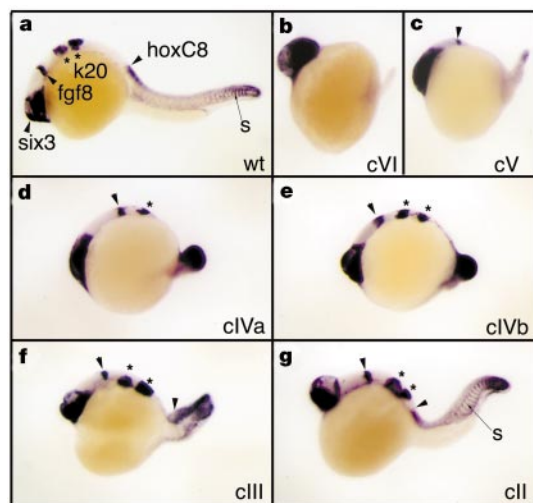


Figure 2 Dose-dependent rescue of posterior territories by *Activin*/Nodal-related factors. **a**, Wild-type embryo (wt) analysed with a combination of probes that label, from anterior to posterior, forebrain (*six3*), the midbrain–hindbrain boundary (*fgf-8*²²), hindbrain rhombomeres 3 and 5 (*Krox20*²³, *k20*, stars), spinal cord (*HoxC8*²⁴) and somites (*s*, *fgf8* and *HoxC8*). **b**, Injection of 1 ng *atv* RNA generates a cVI phenotype; only *six3* expression is observed. Co-injection of 1 ng *atv* RNA with increasing doses of *activin*/NRs RNA rescues structures along the A–P axis (see Table 1) in the following order: midbrain (**c**); midbrain and hindbrain up to rhombomere 3 (**d**), then up to rhombomere 5 (**e**); the whole neurectoderm (**f**); neurectoderm and paraxial somitic mesoderm (**g**).

because marginal cells express high levels of NRs^{7–9}, we propose that induction of this ectopic hindbrain was the result of the secretion of NRs by these grafted cells. Thus, the animal pole territory, although normally not responsive to Activin/Nodal signalling because of its physical separation from the source, can be affected by the local application of these factors, showing that the receptors and required elements of the downstream signalling pathway are present.

We have shown that the zebrafish embryo is patterned along the animal–vegetal axis by the morphogenetic activity of TGF β factors belonging to the Activin/NRs subgroup, with induction of the anterior-most mesendodermal fate associated with the highest Activin/NRs activity and development of anterior-most ectoderm independent of this signalling pathway. Initially, the A–P neural axis in vertebrates was thought to depend on vertical induction through signals from the dorsal mesoderm, which underlies the neur ecto-

derm during gastrulation¹⁸. Using planar explants of dorsal mesoderm and ectoderm, however, it was shown that A–P neural pattern can be induced in ectoderm by planar signals emanating from the mesoderm^{19,20}. We have shown here that A–P neur ectodermal patterning is established in the absence of mesendoderm, and we propose that Activin/NRs are the signals responsible for planar induction. In addition to the currently accepted paradigm for regulated D–V patterning by the interaction between ventral bone morphogenetic proteins factors and their dorsal inhibitors Noggin and Chordin, we propose that patterning along the A–P axis is generated by inter-regulation between another class of TGF β —the Activin/NRs and their antagonist, Activin. □

Methods

In situ hybridization

Whole-mount *in situ* hybridization was done as described²¹. SH372, six3, pax2.1 and vHNF1 clones used to generate RNA antisense probes were isolated during a large-scale *in situ* hybridization screen intended to characterize genes whose expression is spatially restricted during embryogenesis (B.T., C.T., unpublished data).

Embryo injections and lineage tracing

RNA injections were done as described⁵. The clones of cells derived from the blastomeres injected with *cyclops*, *squint* or *activin* RNA were monitored by epifluorescent detection of GFP used as a lineage tracer, translated from co-injected RNA (100 pg). Control embryos received GFP RNA at the same doses and did not give any phenotype.

Received 9 July; accepted 30 November 1999.

- Hogan, B. L. M. Bone morphogenetic proteins: multifunctional regulators of vertebrate development. *Genes Dev.* **10**, 1580–1594 (1995).
- Zimmerman, L. B., De Jesus-Escobar, J. M. & Harland, R. M. The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4. *Cell* **86**, 599–606 (1996).
- Piccolo, S., Sasai, Y., Lu, B. & De Robertis, E. M. Dorsoroventral patterning in *Xenopus*: inhibition of ventral signals by direct binding of chordin to BMP-4. *Cell* **86**, 589–598 (1996).
- Sasai, Y., Lu, B., Steinbeisser, H. & De Robertis, E. M. Regulation of neural induction by the chd and Bmp-4 antagonistic patterning signals in *Xenopus*. *Nature* **376**, 333–336 (1995).
- Thisse, C. & Thisse, B. Antivin, a novel and divergent member of the TGF β superfamily regulates mesoderm induction. *Development* **126**, 229–240 (1999).
- Thomsen, G. *et al.* Activins are expressed early in *Xenopus* embryogenesis and can induce axial mesoderm and anterior structures. *Cell* **63**, 485–493 (1990).
- Rebagliati, M. R. *et al.* *cyclops* encodes a nodal-related factor involved in midline signaling. *Proc. Natl Acad. Sci. USA* **95**, 9932–9937 (1998).
- Sampath, K. *et al.* Induction of the zebrafish ventral brain and floor plate requires *cyclops/nodal* signalling. *Nature* **395**, 185–189 (1998).
- Feldman, B. *et al.* Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* **395**, 181–185 (1998).
- Kimmel, C. B., Warga, R. M. & Schilling, T. F. Origin and organization of the zebrafish fate map. *Development* **108**, 581–594 (1990).
- Woo, K. & Fraser, S. C. Order and coherence in the fate map of the zebrafish nervous system. *Development* **121**, 2595–2609 (1995).
- Green, J. B. A., New, H. J. & Smith, J. C. Responses of embryonic *Xenopus* cells to activin and FGF are separated by multiple dose thresholds and correspond to distinct axes of the mesoderm. *Cell* **71**, 731–739 (1992).
- Thisse, C. *et al.* *goosecoid* expression in neur ectoderm and mesendoderm is disrupted in zebrafish *cyclops* gastrulas. *Dev. Biol.* **164**, 420–429 (1994).
- Melby, A. E. *et al.* Specification of cell fates at the dorsal margin of the zebrafish gastrula. *Development* **122**, 2225–2237 (1996).
- Gristman, K. *et al.* The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. *Cell* **97**, 121–132 (1999).
- Cheng, A. M. S., Thisse, B., Thisse, C. & Wright, C. E. V. The lefty-related factor *Xatv* acts as a feedback inhibitor of Nodal signaling in mesoderm induction and L–R axis development in *Xenopus*. *Development* (in the press).
- Woo, K. & Fraser, S. E. Specification of the zebrafish nervous system by nonaxial signals. *Science* **277**, 254–257 (1997).
- Spemann, H. *Embryonic Development and Induction* (Yale Univ. Press, New Haven, 1938).
- Ruiz y Altaba, A. Neural expression of the *Xenopus* homeobox gene: *Xhox3*: evidence for a patterning neural signal that spreads through the ectoderm. *Development* **108**, 595–604 (1990).
- Doniach, T. *et al.* Planar induction of anteroposterior pattern in the developing central nervous system of *Xenopus laevis*. *Science* **257**, 542–545 (1992).
- Thisse, C. *et al.* Structure of the zebrafish *snail1* gene and its expression in wild-type, *spadetail*, and *no tail* mutant embryos. *Development* **119**, 1203–1215 (1993).
- Reifers, F. *et al.* *fgf8* is mutated in zebrafish *acerebellar* (ace) mutants and is required for maintenance of midbrain–hindbrain boundary development and somitogenesis. *Development* **125**, 2381–2395 (1998).
- Oxtoby, E. & Jowett, T. Cloning of the zebrafish *Krox-20* gene (*Krx-20*) and its expression during hindbrain development. *Nucleic Acids Res.* **21**, 1087–1095 (1993).
- Prince, V. E., Price, A. L. & Ho, R. K. Hox gene expression reveals regionalization along the anteroposterior axis of the zebrafish notochord. *Development* **208**, 517–522 (1998).
- Schulte-Merker, S. *et al.* The protein product of the zebrafish homologue of the mouse *T* gene is expressed in nuclei of the germ-ring and the notochord of the early embryo. *Development* **116**, 1021–1023 (1992).

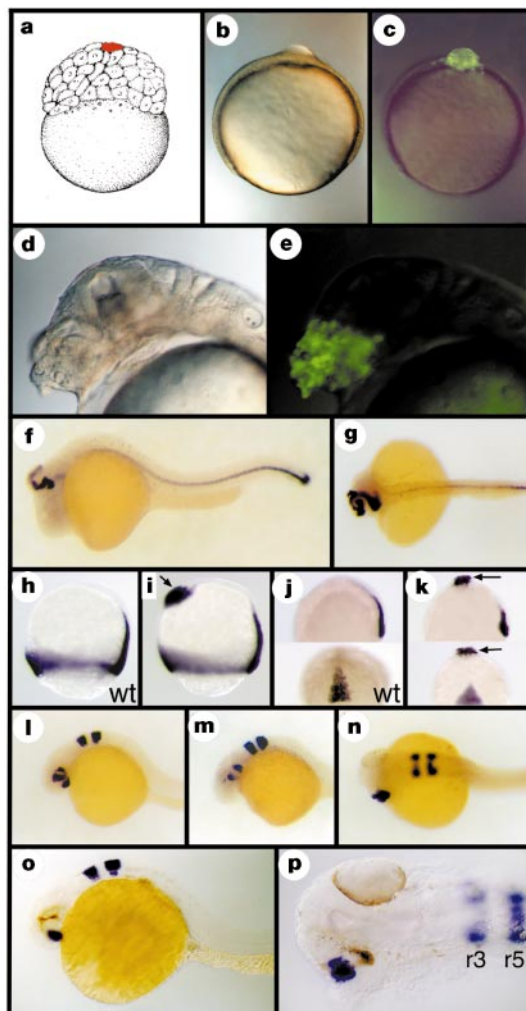


Figure 3 Re-specification of the animal pole region into mesodermal or ectodermal structures by Nodal-related factors. **a**, Diagram of localized *znr1* RNA injection into 128-cell embryos. **b**, *znr1* misexpression leads to the formation of protrusions that are confirmed to contain GFP. **c**, **d–g**, Injection of 30 fg *znr1* RNA induced an ectopic notochord in place of the forebrain of a 24-hour embryo (for 55 out of 57 embryos which had a GFP-containing clone at the animal pole), which was confirmed by detection of *notail*²⁶ (**f,g**). **h,i**, *notail* expression was already detectable at 75% epiboly (compare **i** and **h**, arrow). **j,k**, Induction of notochord was confirmed using a marker of presumptive notochordal territory, *shh*²⁶, at early gastrula stage (60% epiboly; compared **j** and **k**, arrows). **l–p**, Different embryos injected with a lower dose of *znr1* RNA (8 fg) show induction of hindbrain structures into the forebrain (for 38 out of 43 embryos which had a GFP-containing clone at the animal pole) as revealed with *Krox20*²³, a marker of hindbrain rhombomere 3 and 5 (r3, r5). wt, wild-type.

26. Krauss, S., *et al.* A functionally conserved homolog of the *Drosophila* segment polarity gene hh is expressed in tissues with polarizing activity in zebrafish embryos. *Cell* **31**, 1431–1444 (1993).

Acknowledgements

We thank V. Heyer and T. Steffan for technical assistance and O. Nkundwa and A. Karmim for care of fish. B.T. and C.T. were supported by funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg, the Association pour la Recherche sur le Cancer and the Ligue Nationale Contre le Cancer. C.V.E.W. was supported by a NIH grant.

Correspondence and requests for materials should be addressed to C. T.

Separable processing of consonants and vowels

Alfonso Caramazza*, Doriana Chialant*, Rita Capasso†‡ & Gabriele Miceli‡

* Cognitive Neuropsychology Laboratory, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, USA

† Istituto di Ricovero e Cura a Carattere Scientifico S. Lucia, Via Ardeatina 306, 00179 Rome, Italy

‡ Department of Neurology, Università Cattolica, Largo A. Gemelli 8, 00168 Rome, Italy

There are two views about the nature of consonants and vowels. One view holds that they are categorically distinct objects that play a fundamental role in the construction of syllables in speech production^{1–3}. The other view is that they are convenient labels for distinguishing between peak (vowel) and non-peak (consonant) parts of a continuous stream of sound that varies in sonority (roughly the degree of openness of the vocal apparatus during speech)^{4–6}, or that they are summary labels for bundles of feature segments^{7,8}. Taking the latter view, consonants and vowels do not have an independent status in language processing. Here we provide evidence for the possible categorical distinction between consonants and vowels in the brain. We report the performance of two Italian-speaking aphasics who show contrasting, selective difficulties in producing vowels and consonants. Their performance in producing individual consonants is independent of the sonority value and feature properties of the consonants. This pattern of results suggests that consonants and vowels are pro-

cessed by distinct neural mechanisms, thereby providing evidence for their independent status in language production.

AS, a right-handed 41-year-old woman, became aphasic due to ischemic cerebrovascular damage. A computerized tomography scan showed lesions in the left parietal and temporal lobes, and a small lesion in the right parietal lobe. IFA, a right-handed 52-year-old woman, suffered a vascular stroke that resulted in damage to the left supramarginal, angular and superior temporal gyri. The two patients showed similar clinical profiles. They showed no visual, auditory, somatosensory, motor or articulatory deficits; their verbal spans were severely reduced (3 forward and 3 backward for AS, and 3 forward and 2 backward for IFA); and their spontaneous speech was fluent but paraphasic. On a series of assessment tests⁹ they scored very well on phonemic discrimination and auditory word-recognition tasks, but performed poorly on oral naming, reading and repetition tasks (ranging from 40% to 70% correct for AS and 50% to 65% correct for IFA). Performance on the latter tasks was characterized by phonological, morphological, lexical and semantic (only in naming) errors. Both patients could be classified as conduction aphasics. A striking feature of their speech-production performance was contrasting patterns of error rates on vowels relative to consonants: AS made many more errors on vowels and IFA made many more errors on consonants^{10–12}.

To document the suspected double dissociation of selective damage in producing consonants and vowels, AS and IFA were asked to repeat large numbers of words. From an initial corpus of 5,079 (46.3% of total) and 2,671 (43.7% of total) errors produced by AS and IFA, respectively, we analysed the substitution errors, as these were the predominant error type (65.51% for AS and 58.37% for IFA; Table 1). The analyses considered only non-word responses involving three or fewer phoneme substitutions ('Simple Phonological' errors in Table 1). The final error corpus analyses included 771 errors for AS and 939 errors for IFA. The other responses included primarily morphological errors, but also lexical substitution errors (tasca (pocket) → tassa (tax)) and other more complex non-word responses (tracceremo (we will trace) → / trokkomano/). These errors were excluded because they could arise from independent causes (in the case of morphological and lexical substitution errors) or were too complex for unambiguous analysis.

Substitution errors were distributed unequally for consonants and vowels for the two patients, but in opposite directions (Table 2). AS produced nearly three times as many errors on vowels as on consonants and IFA produced nearly five times as many errors on consonants as on vowels. The contrast in performance between the two patients is clearest when we consider their percentage of errors for each phoneme position in words of the same consonant/vowel

Table 1 Error distribution (number of errors and percentage error)

Error type	Examples	AS		IFA	
		N	%	N	%
Substitutions	salire (to climb) → savite	3,327	65.51	1,559	58.37
Deletions	salire → sare	287	5.65	381	14.26
Insertions	salire → savillire	295	5.81	386	14.45
Other types	salire → salte, sarile	645	12.70	182	6.81
Fragments	salire → fa...	446	8.78	146	5.47
No responses		79	1.56	17	0.64
Total errors		5,079	100.00	2,671	100.00
Substitution errors by response type	Examples	AS		IFA	
		N	%	N	%
Simple phonological	salire → solire, sotive, etc.	771	23.17	939	60.23
Complex phonological	salire → tolote, satuvo, etc.	450	13.53	212	13.60
Targets with glide phonological	olio (oil) → omio	109	3.28	93	5.97
Inflections and other morphological	salire → salito (climbed)	1,833	55.09	246	15.78
Lexical	salire → solare (of the sun)	150	4.51	69	4.43
Stress	salire → salevo	14	0.42	0	0.00
Total errors		3,327	100.00	1,559	100.00

Error type: classification into main error types of the total errors for the two patients AS and IFA. Substitution errors by response type: gives the breakdown of substitution errors: simple phonological, involving ≤3 phonemes; complex phonological, involving >3 phonemes; targets with glide phonological, affecting a target word containing a glide (these errors are scored separately because of the uncertain phonological status of glides); inflections and other morphological, resulting in a word morphologically related to the target; lexical, resulting in another word; stress, involving a stress shift in addition to a segmental substitution. Only simple phonological errors were included in further analyses.

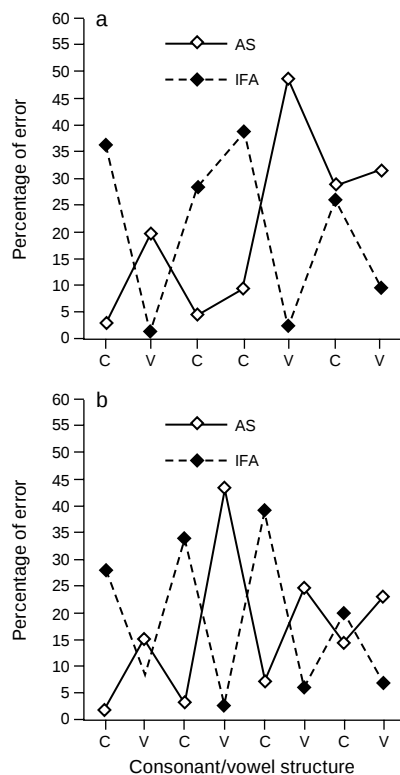


Figure 1 Performance in repetition of a set of words seven or eight phonemes in length. The figure illustrates the complementary performance of the two patients in repetition of items matched for length in number of phonemes and syllable structure. **a**, The data reported refer to items like 'pastore' (shepherd); **b**, the data reported refer to items like 'minatore' (miner).

(C/V) structure. As shown in Fig. 1, the percentage errors on consonants and vowels produced by AS and IFA for words of the same C/V structure are almost in complementary distribution.

Although these results establish that AS and IFA show contrasting selective impairments for vowels and consonants, they are not sufficient to establish that the two types of sounds are categorically distinct objects. It could be argued that damage to a mechanism that is responsible for processing the more sonorous (AS) or less sonorous (IFA) sounds would result in the observed double dissociation. However, this view also predicts that the percentage of errors within the consonant class should vary as a function of sonority¹³: AS should make more errors on the more sonorous than the less sonorous consonants and IFA should show the opposite pattern. We tested this prediction by correlating the percentage of errors for individual consonants with their respective sonority indices⁵. For this and the following analysis we considered only

Table 2 Error distribution by segmental position

		Number of segment errors	Number of segments in words	Error rate
AS	Vowel	737	2,736	0.269
	Consonant	318	3,434	0.093
	Total	1,055		
IFA	Vowel	181	3,397	0.053
	Consonant	1,173	4,159	0.282
	Total	1,354		

Analyses of 771 responses for AS and 939 responses for IFA. Number of segment errors, individual phoneme substitutions produced by each patient. Number of segments in words, occurrences of consonants and vowels in the error corpus. Error rate, ratio of segment errors to the number of segments in the words. The proportions of anticipation and perseveration errors for vowels (AS: 54%; IFA: 57%) and consonants (AS: 30%; IFA: 37%) were comparable for the two patients. Analysis of the substituted consonants and vowels revealed similar patterns for the two patients. In both cases, substitutions which were phonologically close (such as /t/ → /d/) and distant (such as /t/ → /s/) from the target response were equally represented.

simple C/V syllables, so as not to bias the opportunity of errors for consonant substitutions. This is necessary because consonants of different sonority values are distributed unequally across positions within syllables. For both patients, the correlations between sonority values and the error rates were near zero (Table 3). Furthermore, AS was no more likely to produce incorrectly a less rather than a more sonorous segment, and IFA was no more likely to produce incorrectly a more rather than a less sonorous segment. The absence of a correlation between the sonority values of consonants and the probability of making an error on a consonant or producing the consonant in error is not because the range of sonority variation within consonants is too narrow. If anything, liquids (/l/, /r/) are much more similar in sonority to vowels than they are to stop consonants (/p/, /t/): liquids are classified as sonorants in phonetic analyses of speech^{2,8}. Thus, these results show that a deficit to a sonority-based processing mechanism is not the cause of the observed double dissociation.

It could also be argued that selective damage to the sets of features that discriminate among vowels and among consonants, respectively, would result in the observed double dissociation. This view would predict that damage to the features that discriminate among vowels (the dorsal features high, low and back) should result not only in an impairment in processing vowels, but also in greater difficulty for those consonants that are distinguished on the basis of those features (/k/, /g/, /l/, /r/)⁶. However, the error rates for the latter consonants did not differ from those for other consonants (7.9% versus 11.7%, and 27.5% versus 28.3% for AS and IFA, respectively).

Three clear facts have emerged from our analysis of the error performance of patients AS and IFA. First, consonants and vowels can be damaged independently of each other. Second, the distribution of error rates for consonants is not a function of their sonority value. Third, the distribution of errors for consonants is not a function of their dependence on features that distinguish among both vowels and consonants. These three facts have clear implications for theories of speech production and their representation of phonological processes in the brain. The fact that consonants and vowels can be damaged independently of each other rules out, as a general cause of such deficits, an explanation based on the argument

Table 3 Sonority effect analyses

Consonant type	Sonority scale	% error for each target consonant type		% of consonant type produced as an error	
		AS	IFA	AS	IFA
r	8	7.65	21.54	9.12	26.60
l	7	11.27	39.29	32.39	35.71
ɹ	7	16.67	16.67	16.67	5.56
m	6	14.97	14.00	14.97	20.67
n	6	11.36	30.09	11.82	22.57
ɲ	6	15.79	33.33	0.00	8.33
s	5	12.99	30.16	40.26	16.67
ʃ	5	0.00	28.57	0.00	0.00
v	4	8.37	28.99	10.23	32.61
z	4	28.57	19.00	19.00	82.80
f	3	4.76	29.03	0.00	22.58
tʃ	3	5.66	24.76	0.00	43.81
ɔ̃	3	4.76	29.11	3.57	16.46
b	2	16.00	37.70	6.00	14.75
d	2	9.32	57.94	10.56	17.06
g	2	14.00	65.12	8.00	6.98
p	1	8.33	17.65	4.17	20.00
t	1	16.00	17.25	6.59	38.00
k	1	2.38	14.62	17.86	33.08
Spearman's rho		0.16	0.01	0.27	-0.12
		ns	ns	ns	ns
Partial correlation		0.14	-0.15	0.32	-0.13
		ns	ns	ns	ns

Two sets of correlations are reported for each patient. One is between sonority value⁵ and the percentage of errors for each target consonant segment (first two columns). The other is between sonority value and the percentage of occurrences of consonant types produced as errors (last two columns). Error percentages are calculated against the total number of occurrences of each consonant in each set. For each set, a partial correlation was computed with the frequency of occurrence of each consonant in the language partialled out. ns, not significant.

that consonants are more difficult to produce than vowels. More importantly, this fact and the fact that error performance depends neither on the sonority value of individual phonemes nor on their feature properties suggest that consonants and vowels are categorically distinct objects at some level of representation even though they are not categorically distinguishable at a phonetic level. This conclusion is consistent with recent experimental work in speech production with neurologically intact speakers, which has shown that phonological encoding operates over segments (consonants and vowels) and not features¹⁴. Evidence consistent with the possibility that consonants and vowels are represented categorically in perception is provided by the results of a study that stimulated the left superior temporal gyrus of patients with implanted subdural electrode arrays¹⁵. Stimulation impaired discrimination of consonants but not vowels. Importantly, the disruptive effect of the stimulation was equal across all consonants tested, independently of their sonority.

The conclusion that consonants and vowels are represented autonomously does not imply that sonority does not play a role in speech production. The sonority gradient plays a crucial role in determining consonant ordering within the onset and coda of a syllable and in determining syllable boundary. Furthermore, sonority has been used to explain various patterns of speech errors in aphasia^{16–18} and it has been shown to play a direct role in the distribution of consonant errors produced by a non-fluent aphasic patient with frontal lobe damage¹³. The contrasting sets of results suggest that sonority and C/V structure information are used at different levels of the speech production process.

Finally, we considered the possible functional motivation for representing consonants and vowels independently and categorically. The distinction between consonants and vowels plays a crucial role in determining the prosodic structure of speech and in the organization of syllables^{1,16,17}. It has been proposed that syllable structure is not stored directly with our knowledge of the sounds of words but is computed 'on-line' during speech^{19,20}. This on-line process is necessary because the domain of syllabification is not the lexical word (where syllable structure might be represented) but the phonological word²¹. As a consequence syllable boundaries often straddle word boundaries—parts of one word are syllabified with parts of an adjacent word. For example, the phrase 'understand it' is syllabified as un-der-stan-dit and not as un-der-stand-it²¹. The independent representation of C/V structure could serve as the basis for this syllabification process by using this information to assign segments to nucleus and non-nucleus positions in a syllable. □

Received 21 October; accepted 12 November 1999.

- McCarthy, J. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry* 12, 373–418 (1981).
- Clements, G. N. & Keyser, S. J. *CV phonology: A Generative Theory of the Syllable* L1 Monograph Series no. 9 MIT Press, Cambridge, MA, 1983).
- Yip, M. Reduplication and C-V skeleton in Chinese secret languages. *Linguistic Inquiry* 13, 637–661 (1982).
- Selkirk, E. in *The Structure of Phonological Representations* (eds van der Hulst, H. & Smith, N.) Part II (Foris, Dordrecht, 1982).
- Goldsmith, J. A. *Autosegmental and Metrical Phonology* (Blackwell, Oxford, 1990).
- MacNeilage, P. F. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* 21, 499–546 (1998).
- Kenstowicz, M. *Phonology in Generative Grammar* (Blackwell, Oxford, 1994).
- Stevens, K. *Acoustic Phonetics* (MIT Press, Cambridge, MA, 1998).
- Miceli, G., Laudanna, A., Burani, C. & Capasso, R. *Batteria per l'Analisi dei Deficit Afasici* (Centro di Promozione e Sviluppo dell' Assistenza Geriatrica, Roma, 1994).
- Romani, C., Granà, A. & Semenza, C. More errors on vowels than consonants: An unusual case of conduction aphasia. *Brain Language* 55, 144–146 (1996).
- Blumstein, S. E. in *Syllables and Segments* (eds Bell, A. & Hooper, J. B.) 189–200 (North Holland, Amsterdam, 1978).
- Canter, G. J., Trost, J. E. & Burns, M. S. Contrasting speech patterns in apraxia of speech and phonemic paraphasia. *Brain Lang.* 24, 204–222 (1985).
- Romani, C. & Calabrese, A. Syllabic constraints in the phonological errors of an aphasic patient. *Brain Language* 64, 83–121 (1998).
- Roelofs, A. Phonological segments and features as planning units in speech production. *Lang. Cogn. Processes* 14, 173–200 (1999).
- Boatman, D., Hall, C., Goldstein, M. H., Lesser, R. & Gordon, B. Neuroperceptual differences in consonant and vowel discrimination: as revealed by direct cortical electrical interference. *Cortex* 33, 83–98 (1997).

- Buckingham, H. The scan-copier mechanism and the positional level of language production: evidence from phonemic paraphasia. *Cogn. Sci.* 10, 195–217 (1986).
- Beland, R., Caplan, D. & Nespoulous, J.-L. The role of abstract phonological representations in word production: evidence from phonemic paraphasias. *J. Neurolinguistics* 5, 125–164 (1990).
- Christman, S. S. Uncovering phonological regularity in neologisms: contributions of sonority theory. *Clinic. Linguist. Phonetics* 6, 219–247 (1992).
- Ramus, F. & Mehler, J. Language identification with suprasegmental cues: a study based on speech resynthesis. *J. Acoust. Soc. Am.* 105, 512–521 (1999).
- Nespor, M. & Vogel, I. *Prosodic Phonology* (Foris, Dordrecht, 1986).
- Levelt, W. J., Roelofs, A. & Meyer, A. S. A theory of lexical access in speech production. *Behav. Brain Sci.* 22, 1–75 (1999).

Acknowledgements

This research was supported by a grant from the NIH. We thank S. Blumstein, A. Costa, P. Jusczyk, J. Mehler, V. Savova, N. Schiller and K. Stevens for helpful suggestions, and K. Domoto-Reilly for her help in the scoring of the data.

Correspondence and requests for materials should be addressed to A.C. (e-mail: caram@wjh.harvard.edu) or G.M. (e-mail: g.miceli@mlink.it).

Noradrenaline in the ventral forebrain is critical for opiate withdrawal-induced aversion

J. M. Delfs, Y. Zhu, J. P. Druhan & G. Aston-Jones

University of Pennsylvania School of Medicine, Department of Psychiatry, VA Medical Center (151), University & Woodland Avenues, Philadelphia, Pennsylvania 19104, USA

Cessation of drug use in chronic opiate abusers produces a severe withdrawal syndrome that is highly aversive, and avoidance of withdrawal or associated stimuli is a major factor contributing to opiate abuse^{1,2}. Increased noradrenaline in the brain has long been implicated in opiate withdrawal³, but it has not been clear which noradrenergic systems are involved. Here we show that micro-injection of β -noradrenergic-receptor antagonists, or of an α 2-receptor agonist, into the bed nucleus of the stria terminalis (BNST) in rats markedly attenuates opiate-withdrawal-induced conditioned place aversion. Immunohistochemical studies revealed that numerous BNST-projecting cells in the A1 and A2 noradrenergic cell groups of the caudal medulla were activated during withdrawal. Lesion of these ascending medullary projections also greatly reduced opiate-withdrawal-induced place aversion, whereas lesion of locus coeruleus noradrenergic projections had no effect on opiate-withdrawal behaviour. We conclude that noradrenergic inputs to the BNST from the caudal medulla are critically involved in the aversiveness of opiate withdrawal.

Opiate withdrawal results in marked hyperactivity of central noradrenergic neurons^{3,4}. There is evidence that increased noradrenaline is involved in various aspects of the withdrawal response⁵, but it has not been determined where and how increased noradrenaline release contributes to the opiate withdrawal syndrome. The bed nucleus of the stria terminalis (BNST), a component of the extended amygdala⁶, has the highest density of noradrenergic inputs in the brain⁶, and is anatomically connected with other brain areas implicated in drug abuse⁷. We therefore hypothesized that the BNST may be an important site for the actions of noradrenaline during opiate withdrawal.

To determine whether noradrenaline afferents to the BNST are stimulated by opiate withdrawal, we injected a retrograde tracer into the BNST and used triple labelling for the tracer, for tyrosine hydroxylase (a marker for noradrenergic neurons), and for Fos-related antigens (FRAs; a marker of cellular activation). Consistent with previous studies⁷, we found numerous retrogradely

labelled tyrosine hydroxylase-immunoreactive neurons in the A1 and A2 cell groups of the caudal medulla (Fig. 1), with fewer also present in the locus coeruleus. Furthermore, as previously reported^{8,9}, we observed a marked increase in FRA expression in tyrosine hydroxylase-containing neurons of these structures following opiate withdrawal.

Quantification of triple-labelled (tracer, tyrosine hydroxylase and FRA) and double-labelled (tracer and tyrosine hydroxylase) neurons in the A1 and A2 regions of morphine-dependent animals confirmed that most of the tyrosine hydroxylase-immunoreactive cells projecting to the BNST expressed FRAs during withdrawal. In morphine-dependent rats that were given naltrexone ($n = 5$), $81.2 \pm 2.5\%$ of the tyrosine hydroxylase-immunoreactive neurons retrogradely labelled from the BNST in the A1 region were FRA-positive (4.9 ± 0.4 triple-labelled neurons per section), whereas, in non-dependent control animals given naltrexone ($n = 5$), only $13.8 \pm 3.0\%$ of the tyrosine hydroxylase-positive BNST afferents were FRA-immunoreactive (0.6 ± 0.1 triple-labelled neurons per section). In the A2 region of morphine-dependent animals, $76.8 \pm 3.5\%$ of the tyrosine hydroxylase-containing neurons retrogradely labelled from the BNST were FRA-positive after naltrexone-precipitated withdrawal (6.0 ± 1.6 triple-labelled neurons per section), whereas, in non-dependent control animals given naltrexone, only $15.8 \pm 2.0\%$ of such cells were FRA-stained (0.7 ± 0.1 triple-labelled neurons per section). Analysis of variance revealed that both the percentages and raw numbers of FRA-positive, noradrenergic BNST afferent neurons were significantly greater in morphine-dependent than in non-dependent animals ($P < 0.001$ each). Most of the brainstem noradrenergic neurons of the A1 and A2 cell groups that project to the BNST are therefore stimulated by opiate withdrawal. Although many tyrosine hydroxylase-positive neurons in the C1 and C2 medullary regions also become FRA-positive after withdrawal^{8,9}, very few neurons in these regions were retrogradely labelled from the BNST.

As well as measuring the classic somatic (physical) signs of withdrawal, we also assessed opiate withdrawal by using a two-chamber place-conditioning procedure to determine aversion induced by withdrawal. Rats avoid an environment previously paired with precipitated opiate withdrawal¹⁰, which provides a

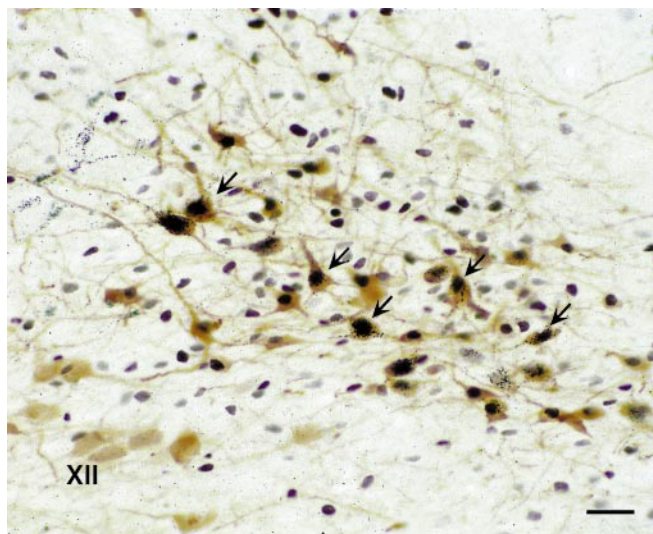


Figure 1 Noradrenergic afferents to the BNST are stimulated by opiate withdrawal. Triple labelling for WGA-gold (small black particles in cell body), FRAs (dark purple–black nuclei) and tyrosine hydroxylase (light brown) in the A2 region of the NTS. Similar results were observed in the A1 and locus coeruleus noradrenergic cell groups, although fewer cells were retrogradely labelled in those areas. Minimal FRA expression was seen in similar sections from control animals (non-dependent saline- and/or naltrexone-injected animals). XII, hypoglossal nucleus. Scale bar, 40 μm .

measure of the negative affective consequences of withdrawal that contribute to aversively motivated drug seeking².

Ascending axons of the A1 and A2 noradrenergic neurons combine to form the ventral noradrenergic bundle (VNAB), whereas those of the locus coeruleus form the dorsal noradrenergic bundle (DNAB). Because many withdrawal-stimulated, retrogradely labelled noradrenergic neurons projecting to the BNST were found in the A1 and A2 cell groups, we tested whether lesions of the VNAB, produced with 6-hydroxydopamine (6-OHDA), would affect the expression of opiate withdrawal behaviours. VNAB lesions markedly reduced withdrawal-induced place aversion in morphine-dependent rats (Fig. 2a), whereas somatic withdrawal symptoms were not significantly affected (Fig. 2b).

To ensure that VNAB lesions did not impair learning ability, we trained a separate group of drug-naïve animals with VNAB lesions in a conditioned place-aversion task by using footshock rather than opiate withdrawal. In contrast to withdrawal-conditioned animals, VNAB-lesioned rats conditioned with footshock had similar aversion scores to unlesioned controls (control, -376.5 ± 56 s; lesion, -371.5 ± 36 s; $P = 0.94$, Student's t -test, $n = 8$ per group). Thus, the VNAB lesion did not generally impair either place learning or aversion, but specifically disrupted withdrawal-induced aversion.

In contrast to the effect of VNAB lesions, destruction of the DNAB with 6-OHDA (Table 1) did not reduce aversive (Fig. 2c) or somatic signs of withdrawal (Fig. 2d). Similarly, lesions of the locus coeruleus noradrenergic system using DSP-4 (50 mg per kg; intraperitoneal; Table 1)¹¹ had no significant effect on withdrawal-induced place aversion (control, -226.4 ± 25 s; lesion, -173.2 ± 38 s; $P = 0.27$, Student's t -test, $n = 8$ per group) or somatic signs (data not shown). Thus, consistent with converging lines of evidence^{12,13}, noradrenergic neurons of the locus coeruleus did not seem to be required for aversive or somatic opiate withdrawal symptoms. Instead, our findings indicate that noradrenergic projections from the A1 and A2 cell groups in the medulla are critically

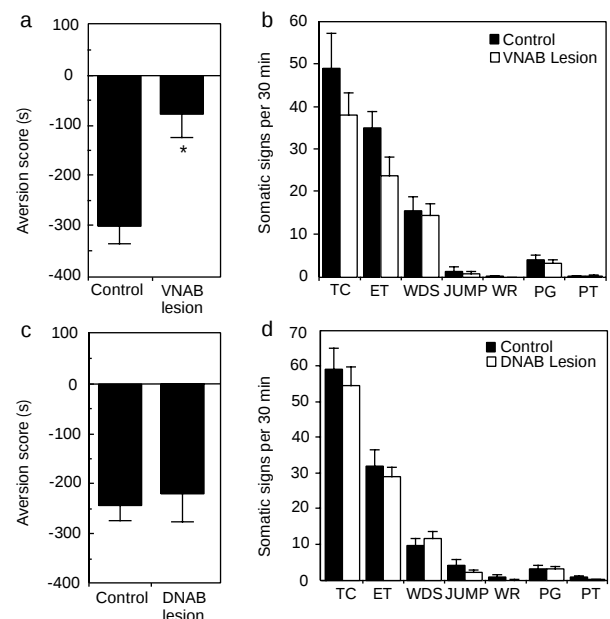


Figure 2 Effects of dorsal (DNAB) and ventral (VNAB) noradrenergic bundle lesions on aversive and somatic signs of opiate withdrawal. **a, c**, Aversion scores are time in the naltrexone-paired side on the test day minus that on the preconditioning day. **b, d**, Number of somatic counts in 30 min. TC, teeth chatter; ET, eye twitch; WDS, wet dog shakes; JUMP, jumping; WR, writhing; PG, penile grooming; PT, paw tremor. Non-opiate-dependent lesioned animals exhibited neither aversion nor somatic signs following naltrexone injection (data not shown). All data are mean $\pm$ s.e.m. ($n = 6$ –8 control animals, $n = 10$ –11 lesioned animals per group). $*P < 0.05$, analysis of variance followed by Fisher's PLSD test for multiple comparisons.

involved in the development of withdrawal-induced place aversion.

We tested the possibility that the BNST is an important site of noradrenaline action during withdrawal by microinjecting noradrenergic drugs directly into it before withdrawal was precipitated. Microinjection of selective β_1 (betaxolol) and β_2 (ICI 118,551) noradrenergic antagonists into the BNST 5 min before naloxone-precipitated withdrawal and place conditioning markedly attenuated conditioned place aversion at 0.1 nmol, and completely eliminated conditioned aversion at 1.0 nmol (Fig. 3a). These effects of the β antagonists were unlikely to result from positive reinforcing effects, as they were administered in both chambers of the conditioning environment. Also, the β -antagonist propranolol has no valence when given alone to morphine-dependent rats¹⁰. Two somatic signs, teeth chattering and eye twitching, were reduced by injection of the β -antagonist cocktail into the BNST (Fig. 3b). Control injections of betaxolol and ICI 118,551 (1.0 nmol) into the overlying striatum (Fig. 4) had no significant effect on aversive (artificial cerebrospinal fluid (ACSF), -220.3 ± 38.5 s; β antagonist, -229.3 ± 44.3 s, $P = 0.88$ Student's t -test, $n = 8$ per group) or somatic signs of withdrawal (data not shown).

To test a second β -receptor antagonist, and to control for a possible local anaesthetic affect of β antagonists, we used isomers of the mixed β_1/β_2 receptor antagonist propranolol. Both isomers have anaesthetic properties, but only the S isomer is a β -adrenoceptor antagonist. Infusion of S -propranolol (5 nmol) in the BNST eliminated withdrawal-induced conditioned place aversion, whereas the same dose of R -propranolol had no effect (Fig. 3c). Similarly, S -propranolol significantly reduced teeth chattering and eye twitching, whereas R -propranolol had no significant effect on any somatic withdrawal sign (Fig. 3d).

Like β antagonists, the α_2 adrenergic agonist clonidine reduces some symptoms of opiate withdrawal in animals and humans^{14–17}. To determine whether α_2 agonists act within the BNST to reduce

opiate withdrawal, we injected a polar analogue of clonidine, ST-91, into this nucleus. Low doses of ST-91 (0.025–0.25 nmol) significantly attenuated conditioned place aversion compared with ACSF-injected controls (Fig. 3e). Clonidine lacks positive motivational effects when given alone to dependent animals¹⁴, indicating that α_2 agonists do not counteract aversion by producing rewarding effects.

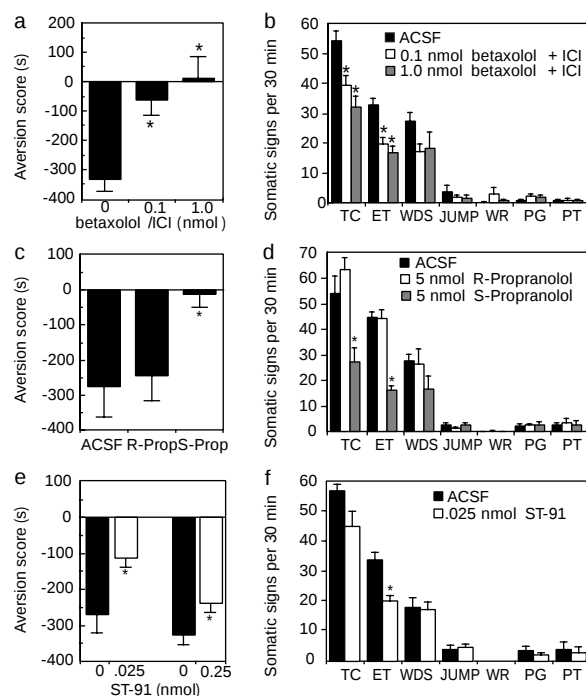


Figure 3 Effects of infusing noradrenergic drugs into the BNST on conditioned place aversion and somatic signs of opiate withdrawal. **a–d**, Effects of the β -antagonist cocktail betaxolol/ICI 118,551 (**a**, **b**) or propranolol isomers (**c**, **d**) on place aversion and somatic signs. **e**, **f**, Effects of ST-91 on place aversion and somatic signs. See Fig. 2 legend for details and abbreviations. All data are expressed as mean $\pm$ s.e.m. ($n = 6–8$ animals per dose). For **a–d**, $*P < 0.05$, analysis of variance followed by Fisher's PLSD for multiple comparisons. For **e**, **f**, $*P < 0.05$, Student's t -test.

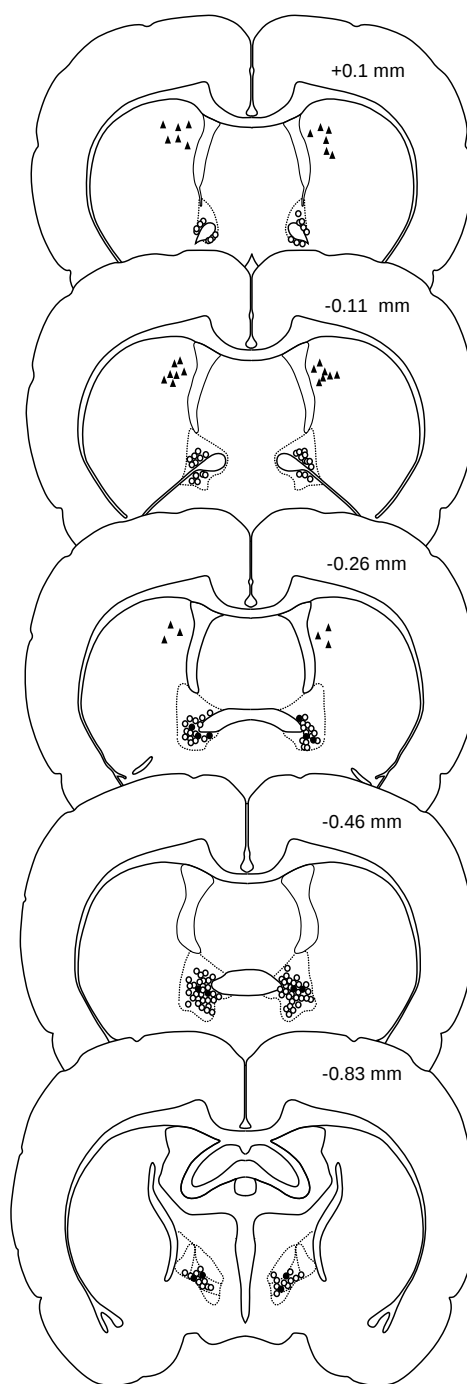


Figure 4 Locations of injection sites. Numbers refer to the distance from the bregma. Effective injections of betaxolol/ICI 118,551, S -propranolol and ST-91 are shown by open circles. The combined aversion scores for these compounds at different levels were, from +0.1 mm to -0.83 mm (mean $\pm$ s.e.m. (n)): -136.4 ± 29 (8); -91.2 ± 90.0 (6); -166.7 ± 25.3 (11); 82.6 ± 40.8 (5); -17.0 ± 68.8 (7). These scores were not significantly different ($P > 0.09$; analysis of variance). The trend for greater effectiveness in mid-caudal BNST may indicate an important role for this sub-area. Ineffective injections are denoted by filled triangles (dorsal injections of betaxolol/ICI 118,551) and filled circles (R -propranolol). Each symbol represents one cannula tip (0.35 mm diameter). The BNST is denoted by the dotted line.

Table 1 Effect of chemical lesions on ^3H -nisoxetine labeled noradrenaline uptake sites in various brain regions

Lesion	Chronic treatment	Binding (%)			
		Cortex	Hippocampus	BNST	Hypothalamus
DNAB (6-OHDA)	Morphine	8.2 $\pm$ 2.9*	5.4 $\pm$ 1.4*	88.7 $\pm$ 8.9	54.0 $\pm$ 7.3*
DNAB (6-OHDA)	Placebo	1.7 $\pm$ 1.0*	6.1 $\pm$ 1.8*	72.9 $\pm$ 0.3	57.0 $\pm$ 9.4*
LC-NE (DSP-4)	Morphine	55.2 $\pm$ 17.0†	33.6 $\pm$ 13.9†	91.3 $\pm$ 6.1	70.1 $\pm$ 9.2†
VNAB (6-OHDA)	Morphine	83.5 $\pm$ 13.2	54.7 $\pm$ 7.1*	12.2 $\pm$ 6.4*	10.7 $\pm$ 3.4*
VNAB (6-OHDA)	Placebo	75.1 $\pm$ 11.4	84.5 $\pm$ 10.7	2.9 $\pm$ 1.4*	9.7 $\pm$ 2.7*

Data represent binding in non-lesioned animals $\pm$ s.e.m. All animals were killed 24 h after the final behavioural test. Note that 6-OHDA lesions of the VNAB markedly reduced noradrenaline uptake sites in the BNST and hypothalamus, with lesser effect on the cortex and hippocampus. DNAB and DSP-4 lesions caused significant reductions in noradrenaline uptake sites in the cortex and hippocampus, with minimal effects on the BNST. LC-NE, locus coeruleus.

* $P < 0.05$, analysis of variance followed by Fisher's PLSD test for multiple comparisons ($n = 6$ –8 control, $n = 10$ –11 lesioned animals per group). † $P < 0.05$, unpaired t -test ($n = 8$ per group).

Reductions in eye twitching and non-significant decreases in teeth chattering were also observed (Fig. 3f).

These results indicate that noradrenergic afferents to BNST are critically involved in the aversive properties of opiate withdrawal. Our anatomical and lesion studies show that these afferents originate in the A2 and A1 cell groups and project to the BNST through the VNAB. Injections of noradrenergic compounds into the BNST indicate that noradrenaline in the BNST promotes withdrawal-induced aversion by stimulating β -adrenergic receptors. This finding reveals a possible site of action for systemic β antagonists that attenuate aversive signs of opiate withdrawal¹⁰. Microinjection of the α 2-autoreceptor agonist ST-91 into the BNST, which reduces the release of noradrenaline, also reduced withdrawal aversion, which is consistent with the idea that noradrenaline release in the BNST during withdrawal is important for withdrawal aversion. Moreover, this result indicates that the BNST is a possible locus for the therapeutic effects of α 2-adrenergic agonists such as clonidine^{14,15}. Our findings also support other evidence that noradrenaline from the locus coeruleus, long believed to be critical for opiate withdrawal³, is not necessary for the associated somatic or aversive responses^{12,13,18,19}.

It is possible that some of the effects reported here resulted from the diffusion of microinjected compounds into neighbouring structures. However, a regional analysis of cannula placements revealed that rostral or caudal injections of adrenergic drugs in the BNST produced no larger effect on place aversion than did centrally located injections (Fig. 4 legend). Also, injections of β antagonists dorsal and equally close to the lateral ventricle (Fig. 4) had no effect on somatic or aversive signs of withdrawal. These results indicate that the BNST is the likely site of action of compounds injected in this study. However, additional experiments are needed to determine whether neighbouring structures, or other parts of the extended amygdala⁵, are also possible sites of noradrenaline action during withdrawal.

The small reduction in teeth chattering and eye twitching following injections of β antagonists into the BNST indicate that noradrenergic projections to the BNST may participate in a subset of somatic withdrawals behaviours. As other somatic signs are attenuated when β antagonists, or α 2 agonists, are administered systematically^{3,10}, it is likely that many of the somatic symptoms of withdrawal involve targets of noradrenaline other than the BNST.

The differential effects of VNAB lesions and BNST infusions on aversive versus somatic symptoms have implications for the relation between these two indices of withdrawal. It has often been assumed that withdrawal aversion is secondary to activation of visceral afferents associated with somatic withdrawal responses. Indeed, evidence that A1 and A2 neurons mediate certain organismic responses to visceral stimuli²⁰ indicates that aversive and somatic signs of withdrawal may depend on similar brainstem processes. However, microinjections of opiate antagonists into the nucleus accumbens (NAC) or amygdala of morphine-dependent rats produce aversive but not somatic signs of withdrawal²¹. Similarly, aversive withdrawal signs can be elicited by doses of naloxone that do not evoke somatic signs²², and they are unaffected by treatments

that completely block the expression of somatic signs¹⁰. These results, together with our findings, indicate that distinct circuits may be responsible for somatic and aversive signs of withdrawal²³, and that the marked reduction in aversion observed after interfering with noradrenaline in the BNST is not dependent on the small decrease in somatic symptoms produced by these manipulations.

The BNST has reciprocal connections with a variety of limbic (NAC and amygdala) and autonomic (periaqueductal gray, hypothalamus and nucleus tractus solitarius structures⁵ that may be involved in withdrawal behaviours¹². Increased noradrenaline during opiate withdrawal has also been linked to decreased dopamine in the NAC^{24,25}, which is important in the opiate withdrawal response²⁶. The location for this apparent interaction between noradrenaline and dopamine during withdrawal is unknown. However, the present findings indicate that the BNST, may be involved through its noradrenaline input and its direct projections to the NAC and midbrain dopaminergic neurons.

From a clinical perspective, avoidance of the aversive aspects of opiate withdrawal is a significant motivating factor for drug seeking in addicts². Our results therefore indicate that the BNST may be an important component of the brain systems involved in the maintenance of drug taking and addictive behaviour. □

Methods

Animal subjects and drug administration

Male Sprague-Dawley rats (Taconic) weighing 250–300 g were used. Morphine dependence was induced by two 75-mg morphine pellets (NIDA) implanted subcutaneously 4–6 days before behavioural testing. Rats were housed in accordance to NIH guidelines on a 12-h light/dark cycle with food and water available ad libitum. All procedures were approved by the institutional animal care and use committee.

Triple immunohistochemistry protocol

Microinjections of a retrograde tracer (wheatgerm agglutinin- α -HRP gold (WGA-gold) or cholera toxin b subunit (CTb)) into the BNST were made as described^{27,28}. Injection sites were centred on the ventral subcommissural division of the BNST with minimal spread into neighbouring structures. Rats were made dependent on morphine, and withdrawal was precipitated 5 days later with naltrexone (1 mg per kg, intraperitoneal). Rats were perfused 2 h after the naltrexone injection and brains were processed for visualization of WGA-gold or CTb, FRAs and tyrosine hydroxylase in the same tissue sections using previously published techniques^{9,27,28}. Quantification was performed by counting the cells within the A1 and A2 regions that were both positive for tyrosine hydroxylase and retrogradely labelled. The number of such double-labelled cells that also contained FRA-positive nuclei was then determined. Cells were counted from every third section to avoid double counting.

6-hydroxydopamine lesions

Rats were anaesthetized with Nembutal (50 mg per kg, intraperitoneal), and bilateral infusion cannulae (30 gauge) were aimed at the dorsal or ventral noradrenergic bundles²⁹. Infusions of 2 μ l of the selective neurotoxin 6-hydroxydopamine or vehicle (DNAB, 2 μ g μ l⁻¹; VNAB, 3 μ g μ l⁻¹; in 0.1% ascorbic acid 0.9% saline) were made over 8 min. Ten days after the surgery, animals were implanted with morphine or placebo pellets and testing began 4 days later. Animals were killed by decapitation 24 h after the final behavioural test and brains were sectioned. All lesions were quantified using ^3H -nisoxetine autoradiography for noradrenaline uptake sites³⁰. Similar lesions have been shown to have no significant effect on dopamine levels in the caudate-putamen²⁹.

Intracerebral microinjections

Under Nembutal anaesthesia, bilateral indwelling guide cannulae (22 gauge) were implanted into the BNST. Cannulae were angled to avoid the lateral ventricle and were

fixed to the skull using dental acrylic. About a week after surgery, animals were implanted with morphine pellets and behavioural testing began 4 days later. Infusions of 0.5 µl per side were made through 28-gauge injector cannulae over 1 min, and cannulae were left in place for 1 min. All drugs were made fresh each day and were dissolved in ACSF. Dye was injected after the experiment to mark the injection site in all animals.

Conditioned place-aversion procedure

A balanced place-conditioning procedure was used to measure aversion in a chamber with two distinct sides¹⁰. On the first day (preconditioning day), rats were allowed free access to both sides of the chamber for 15 min. Animals that spent more than 80% of the time on one side were eliminated. On the next two days (pairing days), the animals were given an intraperitoneal injection of naltrexone (1 mg per kg) or saline, and were confined to one side for 30 min. Animals given naltrexone on pairing day 1 were given saline on pairing day 2 and confined to the opposite side, and vice versa. All adrenergic drugs were micro-injected on each of the two pairing days 5 min before naltrexone or saline; controls were similarly injected with ACSF. During pairing, an observer scored each occurrence of somatic withdrawal signs. On day 4 (test day), animals were given no drug injections and were returned to the test apparatus for 15 min with free access to both compartments, and the time spent in each compartment was measured. For shock training, place conditioning was carried out in drug-naïve animals as above, except that, on pairing day 1, animals received a 0.8 mA foot shock (randomly given for 1 s every 3 min through the chamber floor) over the course of the 30-min session; on pairing day 2, they received no foot shock and were confined to the opposite side.

Received 4 October; accepted 9 November 1999.

1. Bechara, A., Nader, K. & van der Kooy, D. A two-separate-motivational-systems hypothesis of opioid addiction. *Pharmacol. Biochem. Behav.* **59**, 1–17 (1998).
2. O'Brien, C. P., Childress, A. R., McLellan, A., Ehrman, R. & Ternes, J. Types of conditioning found in drug-dependent humans. *NIDA Res. Monogr.* **84**, 44–61 (1988).
3. Maldonado, R. Participation of noradrenergic pathways in the expression of opiate withdrawal: biochemical and pharmacological evidence. *Neurosci. Biobehav. Rev.* **21**, 91–104 (1997).
4. Aston-Jones, G., Shiekhattar, R., Rajkowski, J., Kubiak, P. & Akaoka, H. in *The Neurobiology of Opiates* (ed. Hammer, R.) 175–202 (CRC Press, Boca Raton, Florida, 1993).
5. Alheid, G. F., de Olmos, J. S. & Beltramini, C. A. in *The Rat Nervous System* 2nd edn (ed. Paxinos, G.) 495–578 (Academic, San Diego, 1995).
6. Brownstein, M. J. & Palkovits, M. in *Handbook of Chemical Neuroanatomy, Vol. 2. Classical Transmitters in the CNS, Part 1* (eds Bjorklund, A. & Hokfelt, T.) 23–54 (Elsevier, Amsterdam, 1984).
7. Terenzi, M. G. & Ingram, C. D. A combined immunocytochemical and retrograde tracing study of noradrenergic connections between the caudal medulla and the bed nuclei of the stria terminalis. *Brain Res.* **672**, 289–297 (1995).
8. Akaoka, H., Zhu, Y., Shipley, M. T. & Aston-Jones, G. Expression of fos protein in central catecholamine neurons during opiate withdrawal. *Soc. Neurosci. Abstr.* **18**, 374 (1992).
9. Stornetta, R. L., Norton, F. E. & Guyenet, P. G. Autonomic areas of rat brain exhibit increased Fos-like immunoreactivity during opiate withdrawal in rats. *Brain Res.* **624**, 19–28 (1993).
10. Harris, G. & Aston-Jones, G. Beta-adrenergic antagonists attenuate somatic and aversive signs of opiate withdrawal. *Neuropsychopharmacology* **9**, 303–311 (1993).
11. Fritschy, J. M., Geffard, M. & Grzanna, R. The response of noradrenergic axons to systematically administered DSP-4 in the rat: an immunohistochemical study using antibodies to noradrenaline and dopamine-β-hydroxylase. *J. Chem. Neuroanat.* **3**, 309–321 (1990).
12. Christie, M. J., Williams, J. T., Osbourne, P. B. & Bellchambers, C. E. Where is the locus of opioid withdrawal? *Trends Pharmacol. Sci.* **18**, 134–140 (1997).
13. Caille, S. et al. Total neurochemical lesion of noradrenergic neurons of the locus ceruleus does not alter either naloxone-precipitated or spontaneous opiate withdrawal nor does it influence ability of clonidine to reverse opiate withdrawal. *J. Pharmacol. Exp. Ther.* **290**, 881–892 (1999).
14. Nader, K. & van der Kooy, D. Clonidine antagonizes the aversive effects of opiate withdrawal and the rewarding effects of morphine only in opiate withdrawn rats. *Behav. Neurosci.* **110**, 389–400 (1996).
15. Gold, M. S., Redmond, D. E. & Kleber, H. D. Clonidine blocks acute opiate-withdrawal symptoms. *Lancet* (ii), 599–602 (1978).
16. Roehrich, H. & Gold, M. S. Propranolol as adjunct to clonidine in opiate detoxification. *Am. J. Psychiat.* **144**, 1099–1100 (1987).
17. Gross, H. J. Narcotic withdrawal symptoms in heroin users treated with propranolol. *Lancet* (ii), 564–566 (1972).
18. Chieng, B. & Christie, M. J. Lesions to terminals of noradrenergic locus coeruleus neurones do not inhibit opiate withdrawal behaviour in rats. *Neurosci. Lett.* **186**, 37–40 (1995).
19. Britton, K. T., Svensson, T., Schwartz, J., Bloom, F. E. & Koob, G. F. Dorsal noradrenergic bundle lesions fail to alter opiate withdrawal or suppression of opiate withdrawal by clonidine. *Life Sci.* **34**, 133–139 (1984).
20. Blessing, W. W. *The Lower Brainstem and Bodily Homeostasis* (Oxford Univ. Press, New York, 1997).
21. Stinus, L., Le Moal, M. & Koob, G. F. Nucleus accumbens and amygdala are possible substrates for the aversive stimulus effects of opiate withdrawal. *Neuroscience* **37**, 767–773 (1990).
22. Schulteis, G., Markou, A., Gold, L. H., Stinus, L. & Koob, G. F. Relative sensitivity to naloxone of multiple indices of opiate withdrawal: a quantitative dose-response analysis. *J. Pharmacol. Exp. Ther.* **271**, 1391–1398 (1994).
23. Higgins, G. A. & Sellers, E. M. Antagonist-precipitated opioid withdrawal in rats: evidence for dissociations between physical and motivational signs. *Pharm. Biochem. Behav.* **48**, 1–8 (1994).
24. Potthos, E., Rada, P., Mark, G. P. & Hoebel, B. G. Dopamine microdialysis in the nucleus accumbens during acute and chronic morphine, naloxone-precipitated withdrawal and clonidine treatment. *Brain Res.* **566**, 348–350 (1991).
25. McKittrick, C. R., Delfs, J. M., Aston-Jones, G. & Abercrombie, E. D. Differential spontaneous and withdrawal-induced efflux of catecholamines in nucleus accumbens shell of morphine-treated rats. *Soc. Neurosci. Abstr.* **25**, 948 (1999).
26. Harris, G. & Aston-Jones, G. Involvement of D2 dopamine receptors in the nucleus accumbens in the opiate withdrawal syndrome. *Nature* **371**, 155–157 (1994).
27. Drolet, G., Van Bockstaele, E. J. & Aston-Jones, G. Robust enkephalin innervation of the locus coeruleus from the rostral medulla. *J. Neurosci.* **12**, 3162–3174 (1992).

28. Chen, S. & Aston-Jones, G. Extensive projections from the midbrain periaqueductal gray to the caudal ventrolateral medulla: A retrograde and anterograde tracing study in the rat. *Neurosci.* **71**, 443–459 (1996).
29. Everitt, B. J., Robbins, T. W., Gaskin, M. & Fray, P. J. The effects of lesions to ascending noradrenergic neurons on discrimination learning and performance in the rat. *Neuroscience* **10**, 397–410 (1983).
30. Tejani-Butt, S. M. 3H-nisoxetine: a radioligand for quantification of norepinephrine uptake sites by autoradiography or by homogenate binding. *J. Pharm. Exp. Ther.* **260**, 427–436 (1992).

Acknowledgements

We thank C. Walters and K. Altomare for technical assistance; G. Stanwood for assistance with autoradiography; and S. Aston-Jones for assistance with figures. This work was supported by the National Institute of Drug Abuse, NIH. ST-91 was generously donated by Boehringer-Ingelheim.

Correspondence and requests for materials should be addressed to G.A.-J. (e-mail: gaj@mail.med.upenn.edu).

Nogo-A is a myelin-associated neurite outgrowth inhibitor and an antigen for monoclonal antibody IN-1

Maio S. Chen*, Andrea B. Huber*, Marjan E. van der Haar*, Marcus Frank, Lisa Schnell, Adrian A. Spillmann, Franziska Christ & Martin E. Schwab

Brain Research Institute, Department of Neuromorphology, University of Zurich and Swiss Federal Institute of Technology Zurich, 8057 Zurich, Switzerland

* These authors contributed equally to this work

The capacity of the adult brain and spinal cord to repair lesions by axonal regeneration or compensatory fibre growth is extremely limited. A monoclonal antibody (IN-1) raised against NI-220/250, a myelin protein that is a potent inhibitor of neurite growth, promoted axonal regeneration and compensatory plasticity following lesions of the central nervous system (CNS) in adult rats^{1–4}. Here we report the cloning of *nogo A*, the rat complementary DNA encoding NI-220/250. The *nogo* gene encodes at least three major protein products (Nogo-A, -B and -C). Recombinant Nogo-A is recognized by monoclonal antibody IN-1, and it inhibits neurite outgrowth from dorsal root ganglia and spreading of 3T3 fibroblasts in an IN-1-sensitive manner. Antibodies against Nogo-A stain CNS myelin and oligodendrocytes and allow dorsal root ganglion neurites to grow on CNS myelin and into optic nerve explants. These data show that Nogo-A is a potent inhibitor of neurite growth and an IN-1 antigen produced by oligodendrocytes, and may allow the generation of new reagents to enhance CNS regeneration and plasticity.

We have purified the bovine homologue of rat NI-250, bNI220, a membrane protein of high relative molecular mass (M_r) with potent neurite outgrowth inhibitory activity that can be neutralized by monoclonal antibody (mAb) IN-1 (ref. 5). Microsequencing resulted in six peptide sequences that were used to screen cDNA libraries and databases. We isolated several cDNA clones, and a rat expressed sequence tag (EST) clone (EST111410; ref. 6) showed high homology to the bNI220 peptide sequences 2, 3 and 5. DNA sequence analysis and northern blotting indicated that three different transcripts (A, B and C) originate from one gene, resulting from both alternative promoter usage and alternative splicing (Fig. 1b and data not shown). This gene was designated *nogo* (Fig. 1a). Its protein products are Nogo-A (1,163 amino acids; database accession number AJ242961), Nogo-B (360 amino acids; AJ242962) and Nogo-C (199 amino acids; AJ242963). As Nogo-A contains all six peptide sequences obtained from purified bNI220 (Fig. 1a), it probably corresponds to rat NI-250. Nogo-A, -B and -C have a common carboxy terminus of 188 amino acids (the common

domain), and Nogo-A and -B share an amino terminus of 172 amino acids. Nogo-A is longer than Nogo-B by 803 amino acids because of alternative splicing. Although there is no in-frame stop codon in the putative 5'-untranslated region unequivocally defining the start codon, the following evidence suggests that the methionine indicated in Fig. 1a is the start codon for Nogo-A and -B. (1) The sequence around this presumed start codon conforms well to the consensus sequence for translation start (GCCGCC A/G CCATGG). (2) No additional upstream sequences were found by both library screening and 5'-rapid amplification of cloned ends (RACE). (3) Eukaryotic recombinant Nogo-A expressed from the above mentioned methionine has an apparent relative molecular mass of about 180,000 ($M_r \sim 180K$), as estimated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE), which is indistinguishable from the intracellular form of endogenous Nogo-A made by rat oligodendrocytes.

None of the Nogo isoforms possesses a hydrophobic stretch of amino acids at its N terminus that could serve as a conventional signal sequence. All three Nogos have two long hydrophobic domains (35 and 36 amino acids) close to the C terminus. Nogo-A has three and Nogo-B one additional short conserved hydrophobic stretches. This is in agreement with the characterization of

bNI220 as an integral membrane protein⁵.

Nogo-A contains seven potential N-glycosylation and multiple O-glycosylation sites. Nogo-A (like Nogo-B and -C) does not contain any motifs of known cell-adhesion molecules, extracellular matrix proteins, axonal guidance molecules or growth cone repulsive molecules. At their C termini, Nogo-A, -B and -C have a double lysine endoplasmic-reticulum-retention signal, a property that is shared with several myelin membrane proteins (for example, PMP-22 (ref. 7) or MAL⁸).

Northern blot analysis with the C-terminal 188-amino-acid 'common' probe indicated that three major *nogo* transcripts (4.6, 2.6 and 1.7 kb) were present in optic nerve, spinal cord and cerebral cortex (Fig. 1c). In dorsal root ganglion (DRG) we mainly detected the two larger transcripts. In sciatic nerve and PC12 cells the major band was the 2.6-kb transcript; however, upon longer exposure a weak 4.6-kb band could also be detected (Fig. 1c). We detected a diffuse, unidentified band of about 1.3 kb at various intensities in many tissues. Northern blot studies also showed that *nogo* expression is not restricted to the nervous system: the *nogo* C transcript (1.7 kb) was expressed at a high level in skeletal muscle (Fig. 1c), and small amounts of *nogo* B and C transcripts were detected in kidney,

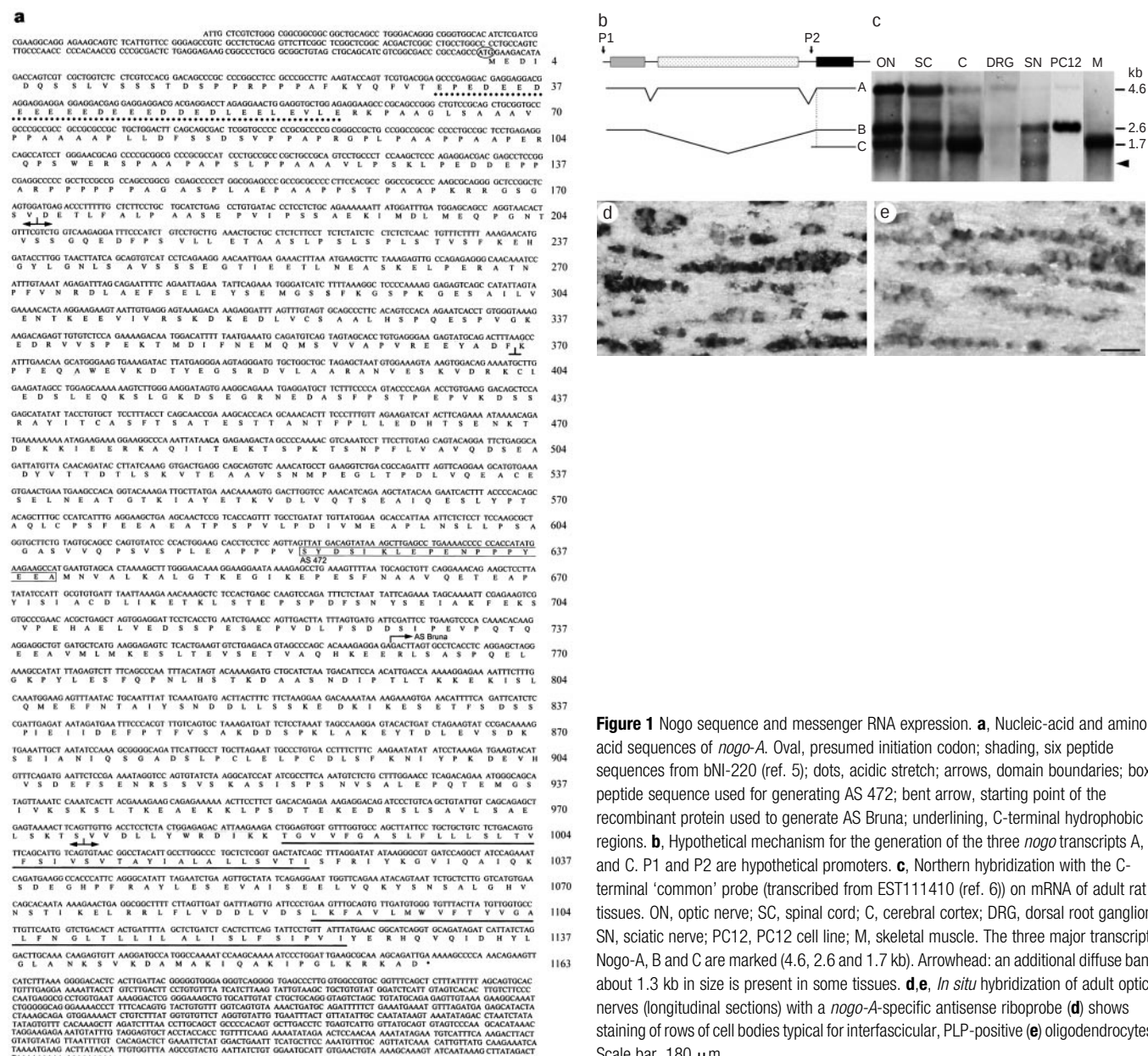


Figure 1 Nogo sequence and messenger RNA expression. **a**, Nucleic-acid and amino-acid sequences of *nogo-A*. Oval, presumed initiation codon; shading, six peptide sequences from bNI-220 (ref. 5); dots, acidic stretch; arrows, domain boundaries; box, peptide sequence used for generating AS 472; bent arrow, starting point of the recombinant protein used to generate AS Bruna; underlining, C-terminal hydrophobic regions. **b**, Hypothetical mechanism for the generation of the three *nogo* transcripts A, B and C. P1 and P2 are hypothetical promoters. **c**, Northern hybridization with the C-terminal 'common' probe (transcribed from EST111410 (ref. 6)) on mRNA of adult rat tissues. ON, optic nerve; SC, spinal cord; C, cerebral cortex; DRG, dorsal root ganglion; SN, sciatic nerve; PC12, PC12 cell line; M, skeletal muscle. The three major transcripts Nogo-A, B and C are marked (4.6, 2.6 and 1.7 kb). Arrowhead: an additional diffuse band about 1.3 kb in size is present in some tissues. **d, e**, *In situ* hybridization of adult optic nerves (longitudinal sections) with a *nogo-A*-specific antisense riboprobe (**d**) shows staining of rows of cell bodies typical for interfascicular, PLP-positive (**e**) oligodendrocytes. Scale bar, 180 μ m.

cartilage, skin, lung and spleen (data not shown). In adult animals the 4.6-kb *nogo A* transcript seemed to be present mainly in the nervous system.

In situ hybridization on adult rat CNS tissue sections using a Nogo-A-specific probe showed labelling of rows of cell bodies in the optic nerve, typical of interfascicular oligodendrocytes (Fig. 1d, e). Besides oligodendrocytes, subtypes of neurons expressed the transcript in brain, spinal cord and peripheral ganglia. Neuronal labelling was particularly conspicuous in the developing nervous system (not shown). Antisera were generated against a synthetic rat Nogo-A-specific peptide (AS 472, Research Genetics) and against a 45K recombinant rat Nogo-A fragment (AS Bruna; Fig. 1a). On western blots, AS 472 and AS Bruna each recognised a protein of about 220K in bovine myelin, q-pool material (myelin extract, enriched in bNI220)⁵ and spinal cord extract (Fig. 2b–d). The same band was recognised by mAb IN-1 (Fig. 2a). Immunohistochemistry of sections of adult rat spinal cord showed identical patterns of staining of CNS white matter by mAb IN-1, AS Bruna and AS 472 (Fig. 2).

To analyse the possible role of Nogo-A in the inhibition of neurite growth by CNS myelin, we used antiserum 472 (Nogo-A specific) and AS Bruna (against all forms of Nogo) in several bioassays. Myelin, myelin protein extract and q-pool enriched in bNI 220 inhibit fibroblast spreading and DRG neurite outgrowth^{1,5,9}. In the presence of AS Bruna, AS 472 or mAb IN-1, the abilities of both bovine spinal cord extract and q-pool to inhibit fibroblast spreading (Fig. 3a,b) and DRG neurite outgrowth (Fig. 3c–e) were greatly decreased. The neutralizing activity of AS 472 could be fully competed by the corresponding peptide 472. A control peptide had no effect.

In a chamber culture system, DRG neurites can grow into sciatic nerve (peripheral nervous system) but not optic nerve (CNS) explants, owing to inhibition by CNS myelin proteins¹⁰. When optic nerves were injected with mAb IN-1, large numbers of ingrowing axons were seen¹. As shown in Fig. 4, pairs of optic

nerves dissected from adult rats were cultured in a chamber, such that DRG neurites had access to one end of each nerve¹⁰. In each culture, one of the two nerves was injected with, and exposed in its side chamber to, AS 472, and the other to pre-immune serum. Nerves injected with pre-immune serum contained no or only a few axons in nine out of ten cultures (Fig. 4b). The few axons, when found, were associated with basement membranes and astrocytes at the surfaces of the nerves. In contrast, nine out of ten optic nerves injected with AS 472 contained considerable numbers of axons, often up to several hundred (Fig. 4c). Many of these axons made contact with myelin (Fig. 4d). These results show that Nogo-A, present in CNS myelin, is a major inhibitory factor for neurite growth.

To test directly for the neurite-growth-inhibiting activity of Nogo-A and its nature as an IN-1 antigen, Nogo-A was expressed as a C-terminal *myc-his*-tagged recombinant protein (recNogo-A) using its own, presumed start site (Fig. 1a). Western blot analysis of cell homogenate with both anti-myc antibody and AS Bruna showed that recNogo-A has an apparent M_r of about 180K on denaturing SDS gels (Fig. 5a, lanes 3, 4). COS cell cultures containing about 15% transfected, recNogo-A-expressing cells were stained after permeabilization with mAb IN-1 and/or AS Bruna. Both antibodies recognized the same transfected cells (Fig. 5b, c). Very similar staining patterns were observed with either of the two antibodies alone (data not shown). We could not detect surface staining of recNogo-A on transfected COS cells.

RecNogo-A produced by a stably transfected Chinese hamster ovary (CHO) cell line (Fig. 5a, lane 3) was tested for its effects on 3T3 fibroblast spreading and DRG neurite outgrowth. Recombinant β -galactosidase isolated from a stable CHO cell line (CHO-LacZ) was enriched in parallel with recNogo-A and was used as a control for inhibitory activity of endogenous CHO proteins in both assays. In the 3T3 fibroblast spreading assay, CHO extract containing recNogo-A (Nogo-A: about 1–5% of total protein) clearly inhibited cell spreading at $10 \mu\text{g cm}^{-2}$ (Fig. 5d). This effect was dose-dependent (data not shown). The inhibitory activity could be

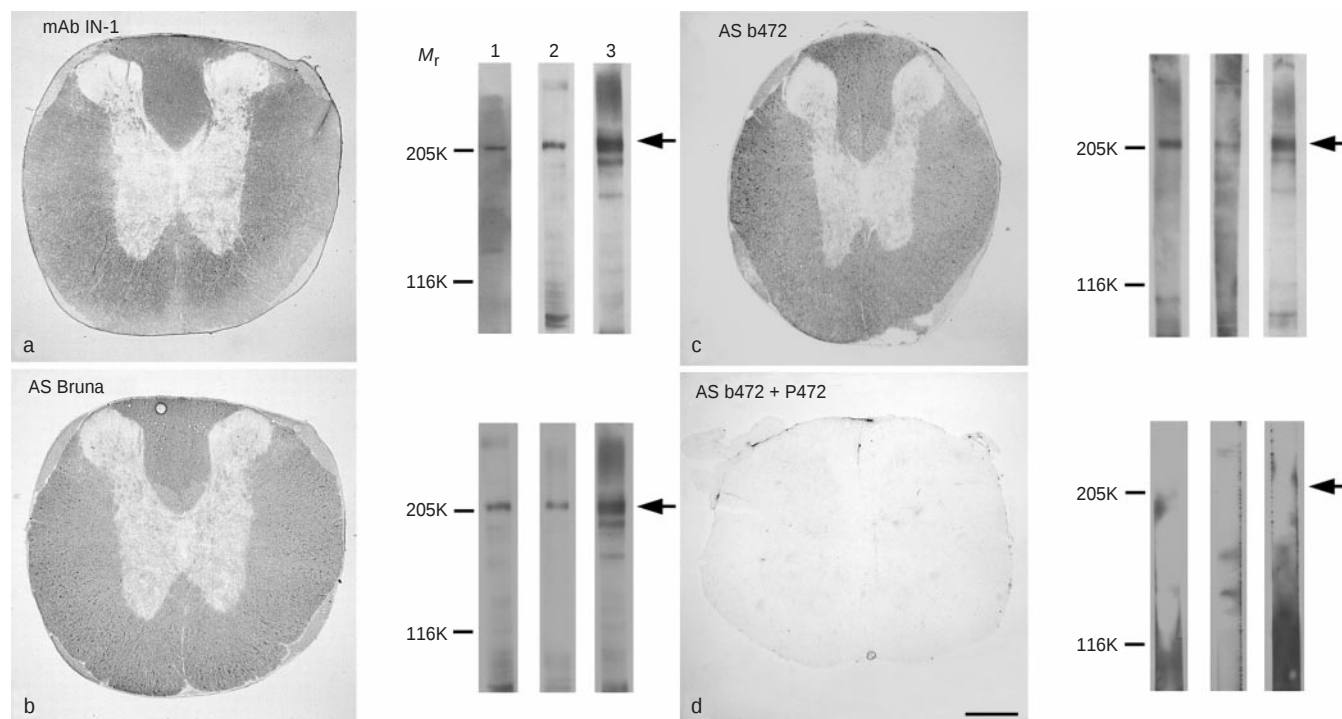


Figure 2 Nogo protein in the adult CNS. Immunohistochemistry on adult rat spinal cord¹⁹ using mAb IN-1, undiluted (**a**); AS Bruna, 1:5,000 (**b**); and affinity-purified AS 472, 1:500 (**c**) showed strong white-matter staining. Pre-incubation of AS 472 with the competing peptide P472 abolished the staining completely (**d**). On western blots all three antibodies

recognised a protein of $M_r \sim 220\text{K}$ (arrow) in myelin (lane 1), spinal cord extract (2) and q-pool enriched in bNI-220 (3). Several lower bands recognized in q-pool by all three antibodies may be breakdown products of the 220K band. Scale bar, 415 μm .

neutralized to background levels by preincubation of the coated protein with mAb IN-1 or AS Bruna, whereas a control antibody (mAb O1) or Bruna pre-immune serum had no effect (Fig. 5d).

Extract of CHO cells containing recNogo-A, but not CHO-LacZ extract, also inhibited neurite outgrowth from primary cultured neurons. Dissociated postnatal day (P)0 rat DRG neurons were inhibited by recNogo-A in a dose-dependent manner (Fig. 5e–g). This inhibitory activity could be neutralized by mAb IN-1, but not by the control mAb O1. Recombinant protein isolated from CHO-LacZ was not inhibitory at 1 and 5 μ g, and addition of mAbs O1 or IN-1 to control cultures had no effect.

Nogo-A has many properties that support the proposal that it is the previously described rat NI-250, a major neurite outgrowth inhibitory protein of CNS myelin and antigen of mAb IN-1. Nogo-A contains all six peptide sequences obtained from purified bNI-220, the bovine equivalent of rat NI-250. It is present in CNS myelin and expressed by oligodendrocytes. In the CNS, Nogo-A is localized to white matter where it is enriched in the inner- and outermost myelin membrane, like myelin-associated glycoprotein (MAG) (confocal and electron microscope immunohistochemistry, data not shown). Antisera against Nogo-A sequences reversed the inhibitory effects of CNS myelin, spinal cord extract enriched in NI-220/250 and cultured adult optic nerve explants, thus reproducing the *in vitro* effects of mAb IN-1. Whether AS 472 and AS Bruna, like mAb IN-1, will promote *in vivo* regeneration of lesioned fibre tracts is currently under investigation.

Recombinant Nogo-A inhibited DRG neurite outgrowth and 3T3 fibroblast spreading in a dose-dependent way and was neutralized by mAb IN-1 or AS Bruna. The relative amount of recNogo-A in the

CHO extract was in the range of 1–5%; thus the amount of recNogo-A effectively coated was 100–500 ng cm^{-2} for 3T3 and 10–50 ng cm^{-2} for DRG assays. This lies in the range of the values that were reported for purified bNI220 (ref. 5). The main inhibitory region is contained in the Nogo-A-specific region, indicating that Nogo-B and -C may not be inhibitory to neurite outgrowth¹¹.

Although the calculated M_r of Nogo-A is 126K, bacterial or COS-cell-expressed Nogo-A, as well as the intracellular pool in oligodendrocytes, has an apparent M_r of about 180K on denaturing SDS gels. Aberrant mobility of proteins on SDS-PAGE has been reported for other highly acidic proteins such as the growth-associated protein GAP-43 and NSP-A¹². The difference to rat myelin NI-250 may be due to glycosylation and the use of slightly different gel-running systems. The Nogo-A sequence is rich in acidic residues and prolines, which are clustered in two regions. Proline-rich regions have structural roles and can be involved in protein–protein interactions. Nogo-A lacks a conventional signal peptide at its N terminus, but preliminary biochemical, immunohistochemical and functional data indicate that Nogo-A may be present on oligodendrocyte cell surfaces, as well as in the endoplasmic reticulum and Golgi, and that this protein can have different topologies. Nogo-A has an endoplasmic-reticulum-retention signal, a property that is shared, however, with several myelin membrane proteins. The possible functional roles of Nogo-A in the endoplasmic reticulum are currently unknown.

Sequence analysis of Nogos has revealed no known motifs of cell surface or matrix proteins involved in axonal guidance; no Ig, fibronectin type III or epidermal growth factor (EGF) domains can be identified, nor is there any homology to known inhibitors of neurite growth such as semaphorins, netrins or ephrins.

Nogos, however, do form a family with a group of proteins, the NSP/s-rex and the RTN-2 proteins, based on the similarity of their C-terminal domains^{12–15}. The sequence similarities of the family members are limited to a region of about 200 amino acids. Outside the homologous region protein sequences are rich in acidic amino acids and, in the case of NSP-A and Nogo-A, rich in prolines. No functions are known for the NSP proteins at present. Some of them

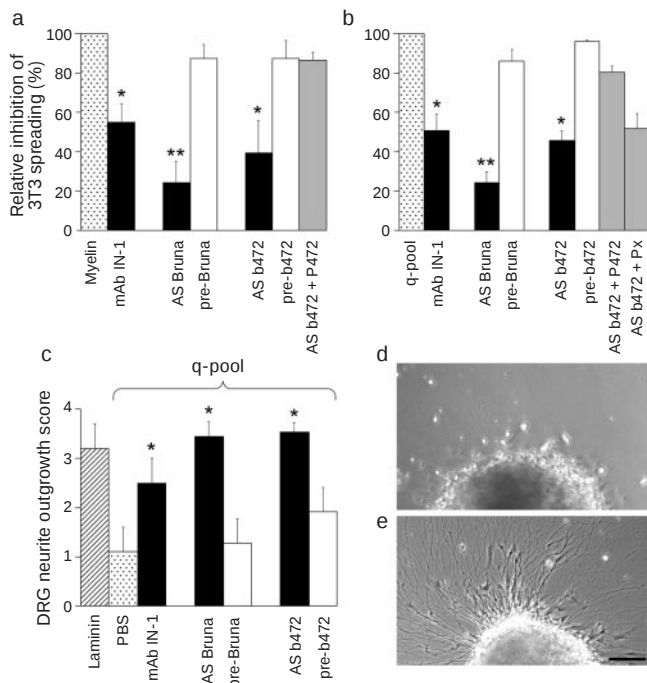


Figure 3 Antisera raised against recombinant Nogo protein neutralize myelin-derived inhibitory activity. 3T3 fibroblasts were plated on cell-culture dishes coated with bovine myelin (a) or q-pool (b). The substrate-coated dishes were pretreated with PBS (maximal inhibition 100%); mAb IN-1, undiluted; AS Bruna, 1:1,000; AS 472, 1:500; or the corresponding pre-immune sera. All three antibodies neutralized the inhibitory effect of both myelin and q-pool. Pre-immune sera had no significant effect. Addition of excess peptide P472 abolished the neutralizing activity of AS 472, whereas the addition of a nonspecific peptide (Px) did not. Similar effects were found for the outgrowth of DRG neurites⁵ (c–e). Outgrowth of DRG neurons was absent on q-pool/laminin-coated dishes (d, score 0); pre-treatment with AS 472 led to abundant outgrowth (e, score 4). Quantifications are shown in (c). Scale bar, 200 μ m. Significance: single asterisk, $P < 0.05$; double asterisk, $P < 0.01$.

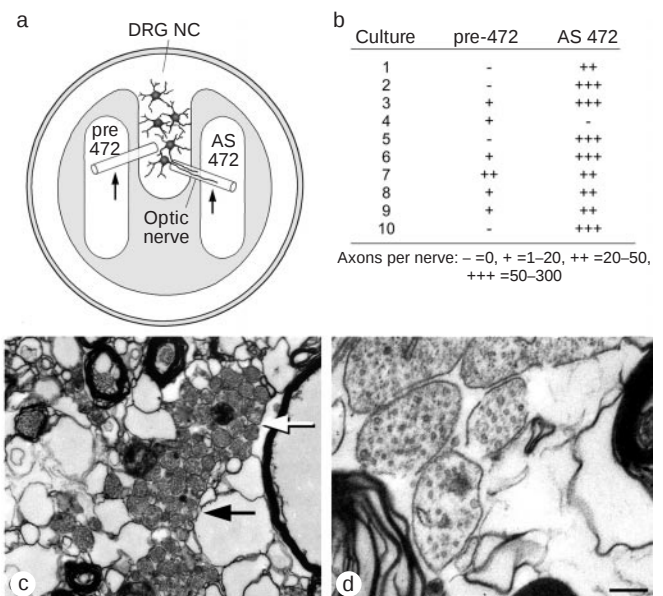


Figure 4 Growth of DRG neurites into adult myelinated optic nerve explants¹⁰ in the presence of AS 472. a, Pairs of adult rat optic nerves were dissected, injected with AS 472 or pre-immune serum and placed into chamber cultures such that one end of the nerves was in contact with dissociated P0 rat DRG neurons. DRG NC, dorsal root ganglion nerve cells. b, Results of systematically screening in the EM for intact axons (arrows in a). c, Regenerated axon bundles (arrows) grow through degenerating AS 472-injected optic nerve. d, Regenerating axons in contact with myelin. Scale bar: c, 0.83 μ m; d, 0.25 μ m.

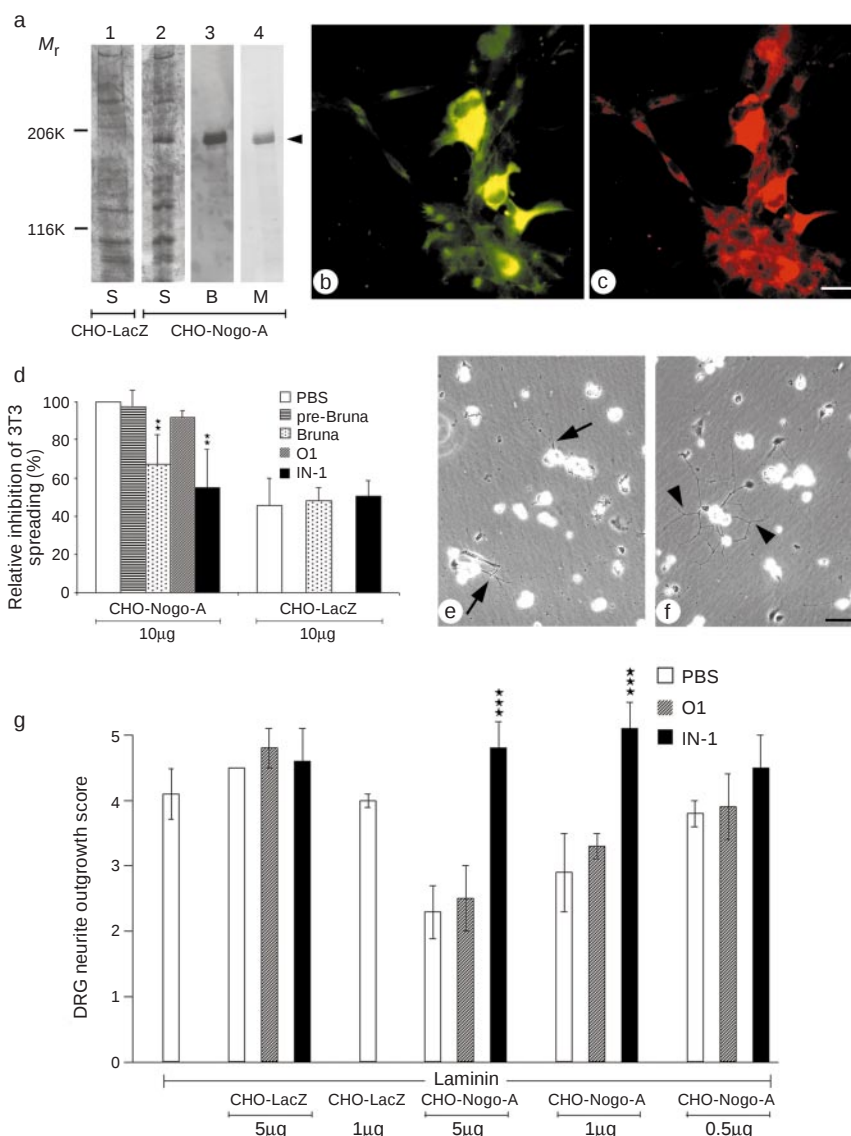


Figure 5 Recombinant Nogo-A is inhibitory and is recognised by mAb IN-1. **a**, Silver-stained gel of myc-his-tagged recLacZ and recNogo-A (lane 1 and 2, respectively; pcDNA3.1myc/his, Invitrogen), enriched over Ni-NTA column (Qiagen) shows Nogo-A at 180K; its identity is confirmed on western blot with AS Bruna (lane 3) and anti-myc antibodies (lane 4). **b,c**, COS cells transiently transfected with Nogo-A were double stained with AS Bruna (**b**) and mAb IN-1 (**c**). Cells stained with AS Bruna were also positive for mAb IN-1. **d**, RecNogo-A-coated dishes were inhibitory to fibroblast spreading; this effect was neutralized by mAb IN-1 or AS Bruna ($P<0.01$). The control IgM mAb O1 or

pre-immune serum were ineffective. CHO-LacZ extract had a partial inhibitory effect, probably due to endogenous CHO proteins, which was not influenced by antibodies. **e-g**, RecNogo-A also inhibited neurite outgrowth of dissociated DRG¹⁰ in a dose-dependent manner (**g**). The activity was neutralized by mAb IN-1 ($P<0.001$), but not by control mAb O1. CHO-LacZ extract was not inhibitory at any of the concentrations used. Examples for scoring are shown in **e** and **f**: **e**, 1 μ g recNogo-A, no or short neurites (arrows), score 2; **f**, 1 μ g recNogo-A pre-incubated with mAb IN-1, long, branched neurites (arrowheads), score 5–6. Scale bar: **b,c**, 46 μ m; **e,f**, 280 μ m.

are predominantly expressed in neurons and several endocrine cell types^{12,14}, with high concentrations in the endoplasmic reticulum (hence their recent renaming as reticulons/RTNs¹⁶).

cDNAs corresponding to rat Nogo-C (vp20; ref. 17), Nogo-B (foocen-m1 (aAAD31020; T. Ito and S. M. Schwarz, unpublished data)) and human Nogo-A (KIAA0886; ref. 18) have been cloned. Orthologues of the C-terminal 200-amino-acid common Nogo/NSP-domain exist in both *Caenorhabditis elegans* and *Drosophila melanogaster*. This indicates that Nogo-A, with its inhibitory activity for nerve regeneration and sprouting, may be a newly evolved relative of an ancient protein of (so far) unknown function.

The neutralization of recNogo-A inhibitory activity by mAb IN-1 and that of myelin inhibitory effects by anti-Nogo-A antiserum show that Nogo-A is the rat neurite growth inhibitor NI-250 (NI-220 in bovine and human). The existence of multiple Nogos and the presence of Nogo-A in cell types other than oligodendrocytes point

to interesting and potentially diverse Nogo functions. New activity-blocking reagents will now become available for functional studies, including studies of the promotion of regeneration and compensatory axon growth following CNS lesions in animals and humans. □

Received 8 June; accepted 17 December 1999.

- Caroni, P. & Schwab, M. E. Antibody against myelin-associated inhibitor of neurite growth neutralizes nonpermissive substrate properties of CNS white matter. *Neuron* **1**, 85–96 (1988).
- Schnell, L. & Schwab, M. E. Axonal regeneration in the rat spinal cord produced by an antibody against myelin-associated neurite growth inhibitors. *Nature* **343**, 269–272 (1990).
- Bregman, B. S. *et al.* Recovery from spinal cord injury mediated by antibodies to neurite growth inhibitors. *Nature* **378**, 498–501 (1995).
- Thallmair, M. *et al.* Neurite growth inhibitors restrict structural plasticity and functional recovery following corticospinal tract lesions. *Nature Neurosci.* **1**, 124–131 (1998).
- Spillmann, A. A., Bandtlow, C. E., Lottspeich, F., Keller, F. & Schwab, M. E. Identification and characterization of a bovine neurite growth inhibitor (bNI-220). *J. Biol. Chem.* **273**, 19283–93 (1998).
- Lee, N. H. *et al.* Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl Acad. Sci. USA* **92**, 8303–8307 (1995).

7. Welcher, A. A., Suter, U., De Leon, M., Snipes, G. J. & Shooter, E. M. A myelin protein is encoded by the homologue of a growth arrest-specific gene. *Proc. Natl. Acad. Sci. USA* **88**, 7195–7199 (1991).
8. Schaeren-Wiemers, N., Valenzuela, D. M., Frank, M. & Schwab, M. E. Characterization of a rat gene, rMAL, encoding a protein with four transmembrane domains in central and peripheral myelin. *J. Neurosci.* **15**, 5753–5764 (1995).
9. Caroni, P. & Schwab, M. E. Two membrane protein fractions from rat central myelin with inhibitory properties for neurite growth and fibroblast spreading. *J. Cell Biol.* **106**, 1281–1288 (1988).
10. Schwab, M. E. & Thoenen, H. Dissociated neurons regenerate into sciatic but not optic nerve explants in culture irrespective of neurotrophic factors. *J. Neurosci.* **5**, 2415–2423 (1985).
11. Chen, M., Schnell, L., van der Haar, M., Oertle, T. & Schwab, M. Inhibitory activity and domain(s) of the myelin protein Nogo-A. *Soc. Neurosci. Abstr.* **25**, 2030 (1999).
12. Roebroek, A. J. *et al.* Cloning and expression of alternative transcripts of a novel neuroendocrine-specific gene and identification of its 135-kDa translational product. *J. Biol. Chem.* **268**, 13439–13447 (1993).
13. Wiczorek, D. F. & Hughes, S. R. Developmentally regulated cDNA expressed exclusively in neural tissue. *Brain Res. Mol. Brain Res.* **10**, 33–41 (1991).
14. Ninkina, N. N., Baka, I. D. & Buchman, V. L. Rat and chicken s-rex/NSP mRNA: nucleotide sequence of main transcripts and expression of splice variants in rat tissues. *Gene* **184**, 205–210 (1997).
15. Geisler, J. G., Stubbs, L. J., Wasserman, W. W. & Mucenski, M. L. Molecular cloning of a novel mouse gene with predominant muscle and neural expression. *Mamm. Genome* **9**, 274–282 (1998).
16. Roebroek, A. J. *et al.* Genomic organization of the human NSP gene, prototype of a novel gene family encoding reticulons. *Genomics* **32**, 191–199 (1996).
17. Morris, N. J., Ross, S. A., Neveu, J. M., Lane, W. S. & Lienhard, G. E. Cloning and preliminary characterization of a 121 kDa protein with multiple predicted C2 domains. *Biochim. Biophys. Acta* **1431**, 525–530 (1999).
18. Nagase, T. *et al.* Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. *DNA Res.* **5**, 169–176 (1998).
19. Rubin, B. P., Dusart, I. & Schwab, M. E. A monoclonal antibody (IN-1) which neutralizes neurite growth inhibitory proteins in the rat CNS recognizes antigens localized in CNS myelin. *J. Neurocytol.* **23**, 209–217 (1994).

Acknowledgements

We thank T. Oertle and C. Bandtlow for suggestions and discussions; T. Flego for technical assistance; R. Schöb and E. Hochreutener for help with the Figures; and C. Brösamle for critically reading the manuscript. This study was supported by grants from the Swiss National Science Foundation, the Maurice E. Muller Foundation (Berne), the International Research Institute for Paraplegia (Zurich) and the Spinal Cord Consortium of the Christopher Reeve Paralysis Foundation (Springfield, New Jersey).

Correspondence and requests for materials should be addressed to M.E.S.

Identification of the Nogo inhibitor of axon regeneration as a Reticulon protein

Tadzia GrandPré^{*†}, Fumio Nakamura^{*†}, Timothy Vartanian[‡] & Stephen M. Strittmatter^{*§}

^{*} Department of Neurology, and [§]Section of Neurobiology, Yale University School of Medicine, New Haven, Connecticut, USA

[‡] Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Institutes of Medicine, Boston, Massachusetts, USA

[†] These authors contributed equally to this work

Adult mammalian axon regeneration is generally successful in the peripheral nervous system (PNS) but is dismally poor in the central nervous system (CNS). However, many classes of CNS axons can extend for long distances in peripheral nerve grafts¹. A comparison of myelin from the CNS and the PNS has revealed that CNS white matter is selectively inhibitory for axonal outgrowth². Several components of CNS white matter, NI35, NI250(Nogo) and MAG, that have inhibitory activity for axon extension have been described^{3–7}. The IN-1 antibody, which recognizes NI35 and NI250(Nogo), allows moderate degrees of axonal regeneration and functional recovery after spinal cord injury^{8,9}. Here we identify Nogo as a member of the Reticulon family, Reticulon 4-A. Nogo is expressed by oligodendrocytes but not by Schwann cells, and associates primarily with the endoplasmic reticulum. A 66-residue luminal/extracellular domain inhibits axonal extension and col-

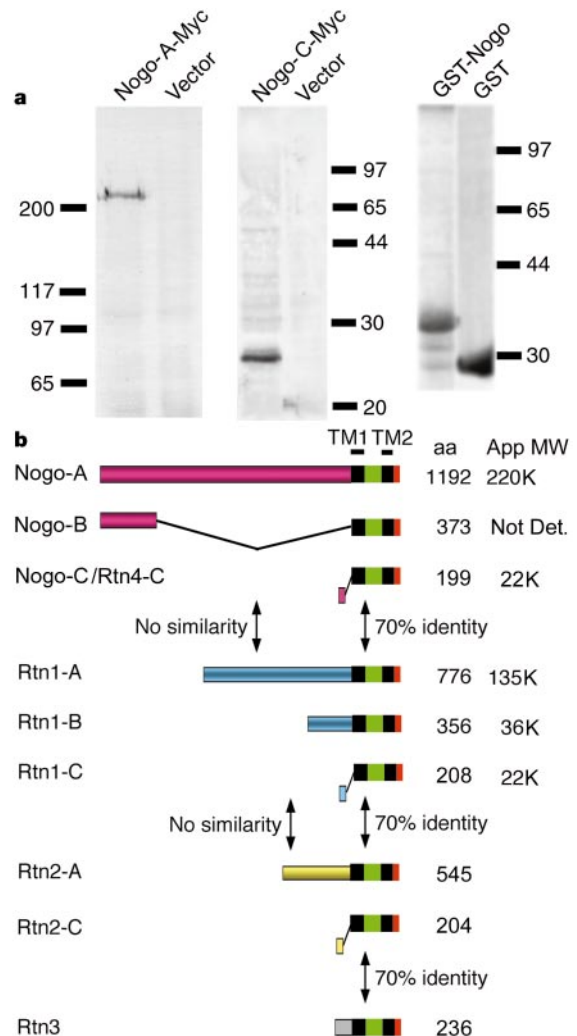


Figure 1 Identification of Nogo. **a**, Membrane extracts from HEK293T cells transfected with vector alone or with an expression vector encoding Nogo-A-Myc or Nogo-C-Myc were analysed by Myc immunoblot (left two panels). The apparent relative molecular mass of Nogo-A-Myc is 225K and that of Nogo-C-Myc is 25K. The purity of the GST-Nogo fusion protein is illustrated by Coomassie Blue staining of SDS-PAGE samples (right panel). **b**, The structure of predicted protein sequences and their relationship to one another. Suffixes A, B and C refer to different splice forms of each family member. The abbreviations are Rtn 1 (Reticulon 1, or neuro-endocrine-specific protein, NSP), Rtn2 (or NSP-like gene 1, NLG1) and Rtn3 (or NSP-like gene 2, NLG2). GenBank accession numbers for the protein sequences: Nogo-A, BAA47909; Nogo-B, AAD31021; Nogo-C, AAF01564; Rtn1-A, A46583; Rtn1-B, I60903; Rtn1-C, I60904; Rtn2-A, 5032055; Rtn2-B, AAC32544; Rtn3, 5174655.

lapses dorsal root ganglion growth cones. In contrast to Nogo, Reticulon 1 and 3 are not expressed by oligodendrocytes, and the 66-residue luminal/extracellular domains from Reticulon 1, 2 and 3 do not inhibit axonal regeneration. These data provide a molecular basis to assess the contribution of Nogo to the failure of axonal regeneration in the adult CNS.

The sequence of six peptides derived from a proteolytic digest of presumed bovine NI250(Nogo) protein has been published³, and the probable full-length complementary DNA sequence for this protein has been deposited in the GenBank. This 4.1-kilobase (kb) human cDNA clone, KIAA0886, is derived from the Kazusa DNA Research Institute effort to sequence random high molecular-weight brain-derived cDNAs¹⁰. This cDNA clone encodes a protein of relative molecular mass 135,000 (M_r 135K) that matches all six of the peptide sequences derived from bovine Nogo.

The predicted molecular mass of protein derived from this clone does not match the observed 200K–250K mobility of Nogo by SDS-polyacrylamide gel electrophoresis (PAGE) (Fig. 1; refs 4, 5). However, the protein has a high number of charged residues that might alter its mobility in gel electrophoresis. Therefore, we measured the electrophoretic mobility of a Myc-epitope tagged version of KIAA0886-derived protein expressed in HEK293T cells. Myc immunoreactivity is detectable with an apparent size in the 225K range under reducing conditions (Fig. 1). Thus, the cDNA directs the expression of a protein with appropriate electrophoretic mobility and the amino-acid sequence to be Nogo. We call this protein hNogo-A (Fig. 1).

The C-terminal third of the hNogo-A sequence shares high homology with the Reticulon (Rtn) protein family (Fig. 1). Reticulon1 has also been called neuro-endocrine specific protein (NSP) because it is expressed exclusively in neuro-endocrine cells¹¹. All Rtn proteins share a 200-amino-acid residue region of sequence similarity at the C terminus of the protein^{11–15}. Related sequences have been recognized in the fly and worm genomes¹⁴. This region is

~70% identical across the Rtn family. Amino-terminal regions are not related to one another and are derived from various alternative RNA splicing events. From analysis of sequences deposited in the GenBank and by homology with published Rtn1 isoforms, we predict three forms of the Nogo protein. Nogo-B (M_r 37K) may possibly correspond to NI35, and explain the antigenic relatedness of the NI35 and NI250 axon outgrowth-inhibiting activity. Nogo-C-Myc has an electrophoretic mobility of 25K by SDS-PAGE and has been described previously as Rtn4-C and vp20 (ref. 15) (Fig. 1).

The conserved C-terminal tail of the Rtn family proteins contains two hydrophobic domains separated by a 66-residue hydrophilic segment (Fig. 2). None of the sequences contains a signal peptide. The predicted topology for these proteins is for the N and C termini to reside in the cytosol, and for the conserved region to associate with the lipid bilayer. For Rtn1-A, experimental evidence shows that the polypeptide behaves as an integral membrane protein, and that the hydrophobic segments of the conserved domain are responsible for this behaviour¹¹. Myc-tagged Nogo is also associated with particulate fractions and is extracted by detergent but not high

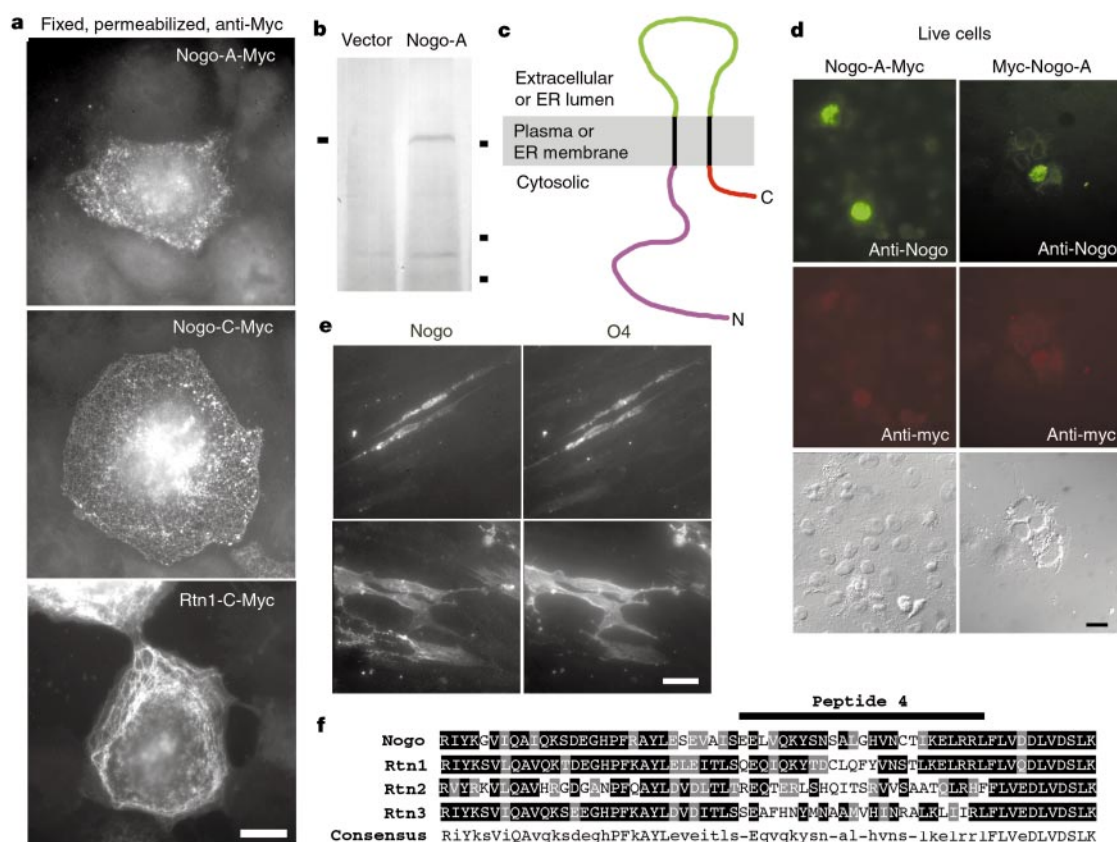


Figure 2 Topology of Nogo protein in transfected COS-7 cells. **a**, COS-7 cells were transfected with plasmids directing the expression of either Nogo-A or Nogo-C or Rtn1-C with a Myc-epitope tag at the C terminus. Thirty-six hours after transfection, cells were fixed and stained for total Myc immunoreactivity after permeabilization. Note the reticular pattern of staining for each of the three proteins. Scale bar, 70 μ m. **b**, HEK293T cells were transfected with control plasmid or a plasmid directing the expression of full-length Nogo-A. Protein was analysed by immunoblot with an antibody generated by immunization of rabbits with the GST–Nogo protein. Nogo is indicated on the left. Markers of 200K, 116K and 97K are shown on the right. **c**, Model of the predicted structure for the Rtn family of proteins. Nogo and Rtns associate with the ER membrane and, to a lesser extent, with the plasma membrane. Green represents the 66-residue predicted luminal/extracellular loop; red, the C terminus; black, the two hydrophobic segments; and purple, the alternatively spliced N-terminal domain. **d**, COS-7 cells prepared as in **a** were incubated with anti-Nogo and anti-Myc primary antibodies before fixation and incubation with secondary fluorescent antibodies. The three panels on the left or on the right are views of the same field showing cell-surface NI-250, cell-surface Myc epitope and

differential interference contrast (DIC) imaging. Note that the epitope recognized by the anti-GST–Nogo antibody is detectable on the cell surface of some cells but that the Myc epitope is not, regardless of whether the epitope is at the N terminus (Myc–Nogo-A) or the C terminus (Nogo–A-Myc) of the protein. The very low level of rhodamine fluorescence in these cells is due to non-specific antibody staining and bleed-through from the fluorescein channel. The percentage of cells exhibiting surface Nogo reactivity was significantly less than the total percentage of transfected cells detected in **a**. The specificity of the anti-Nogo staining was verified by the total absence of staining in vector-transfected cells and by blockade of staining in the presence of GST–Nogo antigen (not shown). Scale bar, 150 μ m. **e**, Unfixed, live chick spinal cord explants were double-stained at 4° with the anti-GST–Nogo antibody (left) and with the O4 oligodendrocyte-specific monoclonal antibody (right). Note the coincident staining of several oligodendrocytes in each pair of panels amongst numerous unstained astrocytes. Anti-Nogo staining was undetectable when GST–Nogo antigen was included in the primary antibody incubation (not shown). Scale bar, 150 μ m. **f**, Amino-acid sequence corresponding to the entire predicted luminal/extracellular loop for each of the proteins shown in Fig. 1.

ionic strength (data not shown). When overexpressed in kidney cells, the Rtn1 protein is localized primarily in the endoplasmic reticulum (ER) in a finely granulated pattern, hence the Reticulon name¹¹. There is a di-lysine ER retention motif at the C terminus of Nogo and most Rtns^{15,16}. In neurons, Rtn1 is expressed throughout processes and is concentrated in growth cones¹⁷. Its localization in transfected kidney cells has led to the suggestion that Rtn1 might regulate protein sorting or other aspects of ER function¹¹. Both the A and C splice forms of Nogo exhibit a reticular distribution when expressed in COS-7 cells, which is similar to that of Rtn1-C (Fig. 2).

The predicted intramembrane topology of the two hydrophobic domains of Nogo suggests that the 66 amino acids between these segments might be localized to the luminal/extracellular face of the membrane (Fig. 2). To explore this further, we generated an antiserum directed against the 66-residue domain (Fig. 2). The antibody detects a low level of surface expression of this epitope, and a Myc epitope at the N or C terminus of expressed Nogo is not detected unless cells are permeabilized (Fig. 2). We attribute this surface staining to a minority of Nogo protein associated with the plasma membrane rather than the ER membrane. These data support a topographic model wherein the N and C termini of the protein reside in the cytoplasm and 66 residues of the protein protrude on the luminal/extracellular side of the ER or plasma membrane.

If Nogo is a major contributor to the axon outgrowth inhibitory characteristics of CNS myelin as opposed to PNS myelin^{4,5,8}, then Nogo should be expressed in adult CNS myelin but not PNS myelin. Northern blot analysis of Nogo expression was carried out using probes derived from the 5' Nogo-A/B-specific region and from the 3' Nogo common region of the cDNA. A single band of ~4.1 kb was detected with the 5' probe in adult rat optic nerve total RNA samples, but not in sciatic nerve samples (Fig. 3). The results indicate that the Nogo-A clone is probably a full-length cDNA and are consistent with Nogo functioning as a CNS-myelin-specific axon outgrowth inhibitor. Northern blot analysis with a 3' probe showed that optic nerve expresses high levels of the Nogo-A messenger RNA and much lower levels of Nogo-B and Nogo-C. Whole brain expresses both Nogo-A and -C, but a number of peripheral tissues (including sciatic nerve) express little or no Nogo (Fig. 3). Nogo-C/Rtn4-C expression has been found in skeletal muscle and adipocytes, as well as in brain¹⁵. Within the Rtn family, optic-nerve expression appears to be selective for Nogo, with no detectable expression of Rtn 1 or Rtn 3 (Fig. 3). Reticulon 2 has not been examined. *In situ* hybridization reveals Nogo mRNA in cells with the morphology of oligodendrocytes in adult rat optic nerve and pyramidal tract (Fig. 3). Within the brain, Nogo expression is also detected in certain neuronal populations. In contrast to Nogo, Rtn1 and Rtn3 are not expressed in optic nerve but mRNA is detected in certain neuronal populations.

We analysed Nogo protein localization in spinal-cord explant cultures treated with platelet-derived growth factor (PDGF) and low serum to induce oligodendrocyte differentiation, using the anti-Nogo antibody and the oligodendrocyte-specific O4 monoclonal antibody. In living cells, both the luminal/extracellular 66-residue loop of Nogo and the O4 antigen are detected on the surface of oligodendrocytes (Fig. 2e). About half the O4-positive cells in these cultures show Nogo surface staining.

The expression of recombinant Nogo in HEK293T cells allows a rigorous test of whether this protein has axon outgrowth-inhibiting effects. Washed membrane fractions from vector- or hNogo-A-Myc-transfected HEK293T cells were added to chick E12 dorsal root ganglion (DRG) explant cultures. Growth cone morphology was assessed after a 30-min incubation at 37° by fixation and rhodamine-phalloidin staining (Fig. 4). The control HEK membranes have no detectable effect on growth cone morphology. The Nogo-A-containing membrane fractions induce the collapse of most DRG growth cones. This growth cone collapse suggests Nogo-A inhibits

axon outgrowth activity, and Nogo inhibition of axon extension can also be shown (see below). The Nogo-C form also exhibits collapse activity, indicating that the shared C terminus of the protein, including the hydrophobic segments and the 66-residue luminal/extracellular domain, contains functionally important residues. Additional inhibitory activity in the N-terminal region of Nogo-A is not excluded by these studies. The sensitivity of more immature explant cultures from E10 chick embryos (Fig. 4) or from E15 rat embryos (data not shown) is substantially less. The developmental regulation of sensitivity is consistent with experiments using partially purified Nogo¹⁸.

In the growth-cone-collapsing Nogo-C protein, the hydrophilic 66-residue luminal/extracellular domain seems more likely than the membrane-embedded hydrophobic domains to interact with the surface of DRG neurons. To test this hypothesis, the 66-residue region of hNogo was expressed in and purified from *Escherichia coli*. Most of the glutathione S-transferase (GST)-Nogo fusion protein accumulates in inclusion bodies, but can be recovered by urea

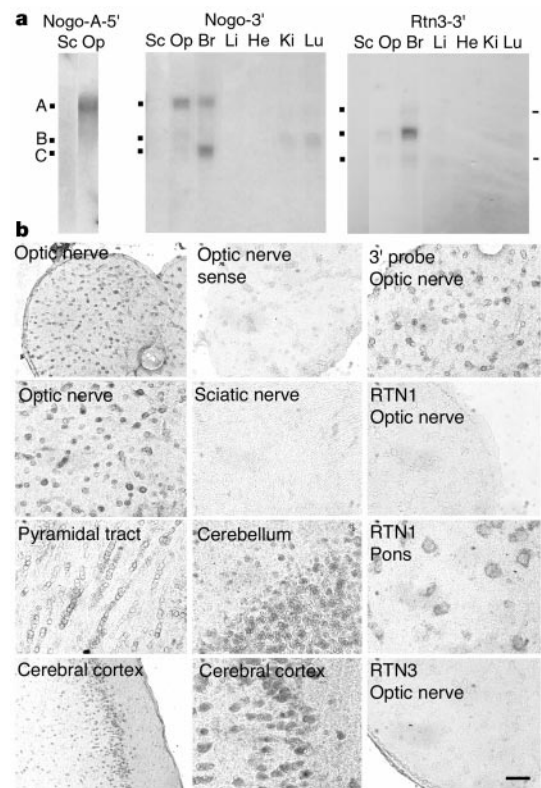


Figure 3 Localization of Nogo mRNA to oligodendrocytes. **a**, Total RNA from the indicated rat tissues was extracted, separated on a formaldehyde-agarose gel and transferred to a nylon membrane. Equal loading was verified by ethidium bromide staining of the 28S and 18S ribosomal RNA (mobility indicated at right). A 5' fragment of hNogo-A complementary to both Nogo-A and -B, or a 3' fragment of hNogo-A complementary to all Nogo mRNA splice forms was used to generate a radioactive probe for Nogo. The migration of the A, B and C splice forms is indicated at left. A similar analysis for the Rtn3 mRNA is shown in the right panel. Sc, sciatic nerve; Op, optic nerve; Br, brain; Li, liver; He, heart; Ki, kidney; Lu, lung. **b**, The distribution of various mRNAs was analysed in adult rat nervous system by *in situ* hybridization with digoxigenin-labelled riboprobes. Left and middle columns, probes complementary to the 5' end of the Nogo gene (splice forms A and B) were hybridized. Top right panel, an antisense probe corresponding to the 3' of the Nogo (splice forms A, B and C) was used. Bottom three panels of the right column, probes complementary to Rtn1C and Rtn 3 mRNAs were used, as indicated. All riboprobes were antisense, except in the panel marked 'sense'. Scale, 200 μ m in all panels except the top and bottom left panels (scale bar, 400 μ m). Note the expression of Nogo in cells with the appearance of oligodendrocytes in the optic nerve and in the pyramidal tract. Rtn1 and Rtn3 are not expressed in optic nerve. Some neurons express Nogo.

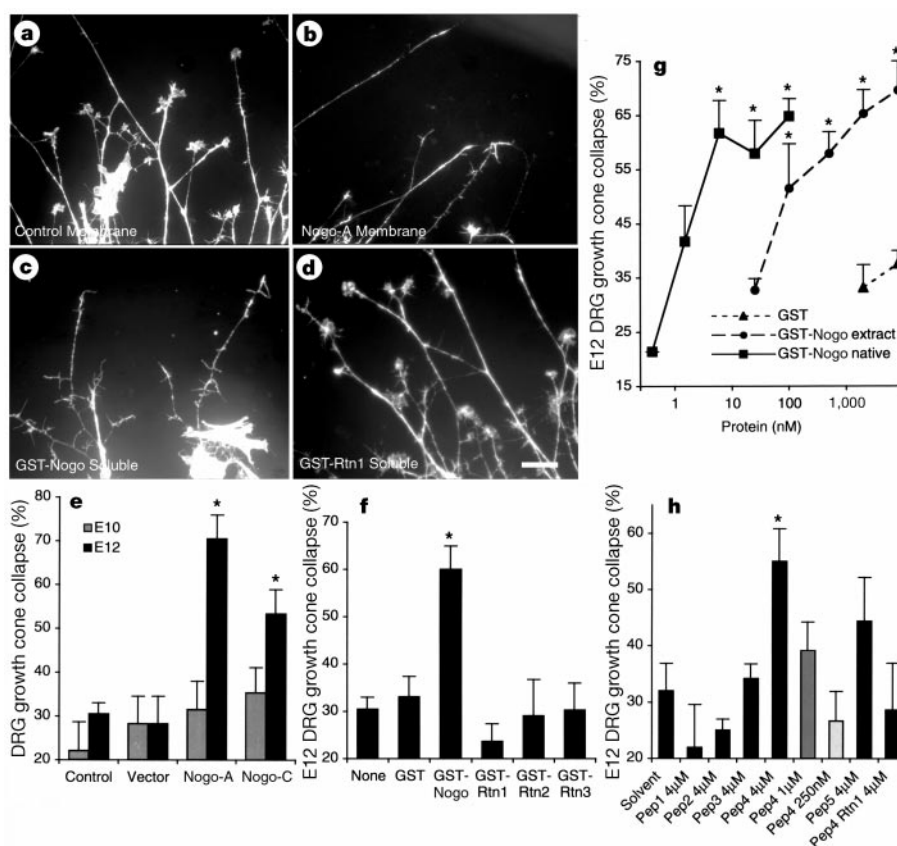


Figure 4 Nogo collapses axonal growth cones. **a–d**, Chick E12 DRG explants were cultured and growth cone collapse was assessed as described in Methods. Cultures were exposed to the following preparations for 30 min before fixation and staining with rhodamine–phalloidin: vector-transfected HEK293T cell membrane fraction at 0.2 mg total protein/ml (**a**); pcDNA3.1-NogoA–Myc vector-transfected HEK293T cell membrane fraction at 0.2 mg total protein ml^{−1} (**b**); 25 nM native GST–Nogo protein (**c**); and 25 nM native GST–Rtn1 protein (**d**). **e–h**, Results from growth cone collapse assays as in **a–d** are quantitated. All results are the means ± s.e.m. from 4–7 determinations.

Membranes were prepared from vector-transfected or Nogo expression plasmid-transfected HEK293T cells and tested at 0.2 mg total protein ml^{−1} with either E10 or E12 DRG explant cultures (**e**). Native GST and GST fusion proteins were tested at 25 nM (**f**), and native and urea extracted fusion proteins were assayed at various concentrations (**g**). Peptides corresponding to portions of the 66-residue luminal/extracellular fragment of Nogo or Rtn1 were tested at the indicated concentrations. Pep1, residues 1–25; Pep2, 11–35; Pep3, 21–45; Pep4, 31–55; Pep5, 42–66 (**h**). Those values significantly different from control are indicated (asterisk, $P < 0.05$, Student's two-tailed t test).

extraction. This restricted region of Nogo possesses potent (median effective concentration (EC_{50}) = 50 nM) growth-cone-collapsing activity for chick E12 DRG neurons (Fig. 4). The urea-extracted protein preparation is likely to present only a small fraction of the Nogo sequence in an active conformation. Therefore, we purified the 10% of GST–Nogo that is soluble in *E. coli* using a glutathione–Sephadex resin. This preparation is even more potent than the urea-extracted protein as a collapsing factor, acutely altering growth cone morphology at concentrations as low as 1 nM. The nanomolar potency is on a par with most known physiologic regulators of axon guidance. Axon outgrowth from DRG neurons and nerve growth factor (NGF)-differentiated PC12 cells is also blocked by this soluble GST–Nogo protein in nanomolar concentrations (Fig. 5). When GST–Nogo is bound to substrate surfaces, axonal outgrowth from DRG neurons or PC12 cells is reduced to undetectable levels (Fig. 5). These are selective effects on axon outgrowth rather than cell survival as GST–Nogo does not reduce the number of neurofilament-positive adherent cells ($137 \pm 24\%$ of GST-treated cultures) nor significantly alter the number of apoptotic nuclei identified by DAPI (4',6-diamidino-2-phenylindole dihydrochloride) staining ($4.0 \pm 1.7\%$ in control cultures and $5.2 \pm 1.1\%$ in GST–Nogo-treated specimens).

Oligodendrocytes appear to express Nogo selectively amongst the Rtns. To explore the selectivity of Nogo's role in the inhibition of axonal regeneration, we considered whether other Rtns exert axon outgrowth-inhibiting activity. The predicted 66-residue luminal/

extracellular fragments of Rtn1, 2 and 3 were expressed as GST fusion proteins and purified in native form. At concentrations in which the Nogo fragment collapses most E12 DRG growth cones, the other Rtns do not alter growth cone morphology (Fig. 4). Thus, the axon regeneration inhibiting activity is specific for Nogo in the Rtn family.

To further define the active domain of Nogo, 25-residue peptides corresponding to segments of the 66-residue sequence were synthesized. The peptide corresponding to residues 31–55 of the extracellular fragment of Nogo exhibits growth-cone-collapsing (Fig. 4) and outgrowth-inhibiting (Fig. 5) activities at concentrations of 4 μM. Although this sequence may provide the core of the inhibitory domain, the 66-residue fragment is clearly required for full potency. Notably, this is the region within the 66-residue domain that shares the least similarity to other Rtn proteins (Fig. 2e), consistent with the other family members being inactive as axon-regeneration inhibitors. Indeed, the 31–55-residue luminal/extracellular peptide of Rtn1 exerts no growth-cone-collapsing activity (Fig. 4).

We have identified Nogo as an oligodendrocyte-specific member of the Rtn family and have shown that a discrete domain of Nogo can inhibit axon outgrowth. Other Rtns do not possess this activity. The expression of Nogo in oligodendrocytes but not Schwann cells may contribute to the failure of axonal regeneration in the adult mammalian CNS as opposed to in the adult PNS. The relative contribution of Nogo as compared with other CNS myelin components to the non-permissive nature of CNS white matter can now

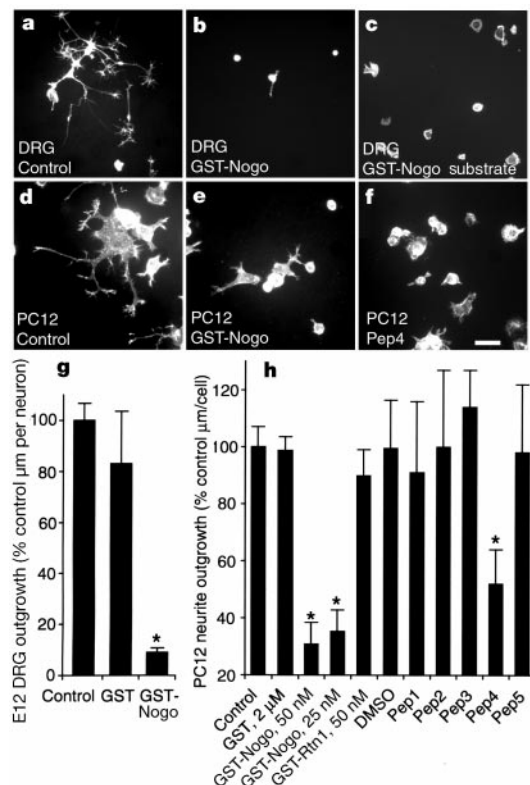


Figure 5 The luminal/extracellular domain of Nogo inhibits neurite outgrowth. **a–f**, Chick E12 dissociated DRG neurons (**a–c**) or NGF-differentiated PC12 cells (**d–f**) were cultured for 10–24 hours in the presence of the indicated proteins or peptides: no added protein (**a**); 100 nM native GST–Nogo protein (**b**); GST–Nogo substrate coating (10 ng protein dried per mm^2) (**c**); 2 μM GST protein (**d**); 25 nM native GST–Nogo protein (**e**); and 4 μM peptide 4 (residues 31–55 of the Nogo luminal/extracellular domain) (**f**). A substrate coating of GST (50 ng protein dried per mm^2) did not alter outgrowth from control levels (data not shown). Neurite outgrowth was visualized by staining with rhodamine–phalloidin. Scale bar, 200 μm . **g–h**, Results from neurite outgrowth assays as in **a–f** are quantitated. For dissociated E12 DRG cultures tested with soluble proteins, the GST concentration was 2 μM and the native GST–Nogo concentration was 100 nM (**g**). For NGF-differentiated PC12 cultures, the protein concentrations are indicated and the peptide concentrations were 4 μM (**h**). The peptide sequences are as in Fig. 4. All results are the means $\pm$ s.e.m. from 4–7 determinations. Those values significantly different from control are indicated (asterisk, $P < 0.05$, Student's two-tailed t test).

be characterized at a molecular level. For example, axonal regeneration in mice lacking Nogo can be compared with that in mice lacking MAG or both MAG and Nogo. IN-1 effects can be studied in animals lacking Nogo expression to determine whether the effect of the IN-1 antibody is due to Nogo blockade. Although the current data are consistent with a role for Nogo in blocking adult CNS axonal regeneration after pathologic injury, this may or may not be related to the physiologic role of Nogo in non-pathologic states. On the basis of localization studies, other Rtns are thought to have a role in ER function¹¹. Most of Nogo is distributed in a reticular pattern in COS-7 cells, and only a minority seems to be accessible at the cell surface. Nogo is found on the surface of some, but not all, oligodendrocytes. The localization of at least part of Nogo's axon outgrowth-inhibiting activity to a discrete region of the protein should facilitate further mechanistic analysis of axon repulsion by this protein. □

Methods

Expression vectors and protein purification

The full-length sequence of Nogo-A (KIAA0886) was generously provided by the Kazusa DNA Research Institute. The full-length coding sequence was amplified by polymerase

chain reaction (PCR) and ligated into the pcDNA3.1–MycHis vector (Invitrogen) to generate a plasmid encoding Nogo-A fused at the C terminus to the Myc epitope (Nogo-A–Myc). Alternatively, the coding sequence was amplified using primers that encode an in-frame Myc epitope immediately N-terminal to the first residue and a stop codon at the C terminus (Myc–Nogo-A). The Nogo-C–MycHis and Rtn1C–MycHis expression vectors were derived in the same fashion except that an adult rat brain cDNA library was used as a template for the PCR reaction with primers based on the Nogo-C or Rtn1C sequences¹¹. These plasmids were transfected into COS-7 or HEK293T by the Lipofectamine (Life Sciences) or the FuGENE 6 (Boehringer Mannheim) method.

A portion of Nogo-A encoding the 66-residue luminal/extracellular fragment of Nogo-A (Fig. 2e) was amplified by PCR and ligated into the pGEX-2T plasmid to yield a prokaryotic expression vector for the GST–Nogo fusion protein. Similar regions of Rtn1, Rtn2 and Rtn3 were amplified by nested PCR using an adult rat brain cDNA library as template and were ligated to pGEX-2T. *E. coli* transformed with these plasmids were induced with IPTG. Soluble, native GST fusion proteins were purified using a glutathione–resin and contained ~75% GST and 25% full-length GST–Nogo or GST–Rtn protein. Most of the GST–Nogo protein was not extractable under non-denaturing conditions, but an 8M-urea extract dialysed against PBS contained over 98% pure GST–Nogo.

Antibody production and immunohistology

Anti-Myc immunoblots and immunohistology with the 9E10 antibody were obtained as described^{19,20}. The GST–Nogo fusion protein was used as an immunogen to generate an anti-Nogo rabbit antiserum. Antibody was affinity purified and used at 3 $\mu\text{g ml}^{-1}$ for immunohistology and 1 $\mu\text{g ml}^{-1}$ for immunoblots. To assess the specificity of the antiserum, staining was conducted in the presence of GST–Nogo protein at 0.1 mg ml^{-1} . For live-cell staining, cells were incubated in primary antibody dilutions at 4° for 1 hour in Hanks balanced salt solution with 0.05% (w/v) BSA and 20 mM Na-HEPES pH 7.3. After fixation, bound antibody was detected by incubation with fluorescently labelled secondary antibodies.

Messenger RNA distribution

For northern blots, total RNA on nylon membranes was hybridized with ³²P-labelled probes generated by the random priming method with the Klenow enzyme as described^{21,22}. The 5' Nogo probe corresponds to the nucleotides encoding residues 1–878, the 3' Nogo probe to the entire Nogo-C coding sequence, and the Rtn3 probe to the entire coding sequence. *In situ* hybridization used the digoxigenin-labelled riboprobe method as described^{21,22}. The 5' Nogo, 3' Nogo and Rtn3 riboprobes correspond to the same sequences used for northern blot probes. The Rtn1C riboprobes are complementary to the entire Rtn1C coding sequence.

Neural culture, growth cone collapse and neurite outgrowth

For the culture of embryonic chick E10 and E12 DRG explants and dissociated neurons, we used methods described for E7 DRG cultures^{19–21,23}. Nerve growth factor-differentiated PC12 cells were cultured as described²⁴. Embryonic spinal cord explants (rat E10 or chick E5) were cultured for 7–14 days in the presence of PDGF-AA to induce differentiation of some cells into mature oligodendrocytes²⁵. The procedure for growth-cone-collapse assays is identical to that for analysis of Sema3A-induced growth cone collapse^{19–21,23}. The method for analysis of total neurite outgrowth has also been described^{21,23,24}. In outgrowth assays, proteins and peptides were added 1 hour after plating to minimize any effect on the total number of adherent cells. To test the effect of substrate-bound GST or GST–Nogo, the protein solutions were dried on poly-L-lysine coated glass, washed and then coated with laminin. For E12 cultures, the neuronal identity of cells was verified by staining with anti-neurofilament antibodies (2H3, Developmental Studies Hybridoma Bank) and neurites were traced by observation of rhodamine–phalloidin staining of F-actin in processes.

Received 12 November; accepted 21 December 1999.

- Benfey, M. & Aguayo, A. G. Extensive elongation of axons from rat brain into peripheral nerve grafts. *Nature* **296**, 150–152 (1982).
- Schwab, M. E. & Thoenen, H. Dissociated neurons regenerate into sciatic but not optic nerve explants in culture irrespective of neurotrophic factors. *J. Neurosci.* **5**, 2415–2423 (1985).
- Wang, L. H. *et al.* In *Neurobiology of Spinal Cord Injury* (eds Kalb, R. G. & Strittmatter, S. M.) 131–153 (Humana, Totowa, New Jersey, 1999).
- Caroni, P. & Schwab, M. E. Two membrane protein fractions from rat central myelin with inhibitory properties for neurite growth and fibroblast spreading. *J. Cell Biol.* **106**, 1281–1288 (1988).
- Spillmann, A. A., Bandtlow, C. E., Lottspeich, F., Keller, F. & Schwab, M. E. Identification and characterization of a bovine neurite growth inhibitor (bNI-220). *J. Biol. Chem.* **273**, 19283–19293 (1998).
- McKerracher, L. *et al.* Identification of myelin-associated glycoprotein as a major myelin-derived inhibitor of neurite growth. *Neuron* **13**, 805–811 (1994).
- Mukhopadhyay, G., Doherty, P., Walsh, F. S., Crocker, R. & Filbin, M. T. A novel role of myelin-associated glycoprotein as an inhibitor of axonal regeneration. *Neuron* **13**, 757–767 (1994).
- Bregman, B. S. *et al.* Recovery from spinal cord injury mediated by antibodies to neurite growth inhibitors. *Nature* **378**, 498–501 (1995).
- Thallmair, M. *et al.* Neurite growth inhibitors restrict plasticity and functional recovery following corticospinal tract lesions. *Nature Neurosci.* **1**, 24–31 (1998).
- Nagase, T. *et al.* Prediction of the coding sequences of unidentified human genes. XII. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*. *DNA Res.* **31**, 355–364 (1998).
- van de Velde, H. J., Roebroek, A. J., Senden, N. H., Ramaekers, F. C. & Van de Ven, W. J. NSP-encoded reticulons, neuroendocrine proteins of a novel gene family associated with membranes of the endoplasmic reticulum. *J. Cell Sci.* **107**, 2403–2416 (1994).

12. Roebroek, A. J. *et al.* Genomic organization of the human NSP gene, prototype of a novel gene family encoding reticulons. *Genomics* **32**, 191–199 (1996).
13. Roebroek, A. J., Contreras, B., Pauli, I. G. & Van de Ven, W. J. cDNA cloning, genomic organization, and expression of the human RTN2 gene, a member of a gene family encoding reticulons. *Genomics* **51**, 98–106 (1998).
14. Moreira, E. F., Jaworski, C. J. & Rodriguez, I. R. Cloning of a novel member of the reticulon gene family (RTN3): gene structure and chromosomal localization to 11q13. *Genomics* **58**, 73–81 (1999).
15. Morris, N. J., Ross, S. A., Neveu, J. M., Lane, W. S. & Lienhard, G. E. Cloning and characterization of a new member of the reticulon family. *Biochim. Biophys. Acta* **1450**, 68–76 (1991).
16. Jackson, M. R., Nilsson, T. & Peterson, P. A. Identification of a consensus motif for retention of transmembrane proteins in the endoplasmic reticulum. *EMBO J.* **9**, 3153–3162 (1991).
17. Senden, N. H. *et al.* Neuroendocrine-specific protein C (NSP-C): subcellular localization and differential expression in relation to NSP-A. *Eur. J. Cell Biol.* **69**, 197–213 (1996).
18. Bandtlow, C. E. & Loschinger, J. Developmental changes in neuronal responsiveness to the CNS myelin-associated neurite growth inhibitor NI-35/250. *Eur. J. Neurosci.* **9**, 2743–2752 (1997).
19. Takahashi, T., Nakamura, F., Jin, Z., Kalb, R. G. & Strittmatter, S. M. Semaphorin A and E act as antagonists of neuropilin-1 and agonists of neuropilin-2 receptors. *Nature Neurosci.* **1**, 487–493 (1998).
20. Takahashi, T. *et al.* Plexin-neuropilin1 complexes form functional semaphorin-3A receptors. *Cell* **99**, 59–69 (1999).
21. Goshima, Y., Nakamura, F., Strittmatter, P. & Strittmatter, S. M. Collapsin-induced growth cone collapse mediated by an intracellular protein related to *unc-33*. *Nature* **376**, 509–514 (1995).
22. Wang, L. H. & Strittmatter, S. M. A family of rat CRMP genes is differentially expressed in the nervous system. *J. Neurosci.* **16**, 6197–6207 (1996).
23. Jin, Z. & Strittmatter, S. M. Rac1 mediates collapsin-induced growth cone collapse. *J. Neurosci.* **17**, 6256–6263 (1997).
24. Strittmatter, S. M., Fishman, M. C. & Zhu, X. P. Activated mutants of the α subunit of G_o promote an increased number of neurites per cell. *J. Neurosci.* **14**, 2327–2338 (1994).
25. Vartanian, T., Fischbach, G. & Miller, R. Failure of spinal cord oligodendrocyte development in mice lacking neuregulin. *Proc. Natl Acad. Sci. USA* **96**, 731–735 (1999).

Acknowledgements

The authors thank A. Fournier for invaluable experimental advice and assistance. This work was supported by grants to S.M.S. from the NIH and the American Paralysis Association, and to F.N. from the Spinal Cord Research Fund of the Paralyzed Veterans of America. S.M.S. is an Investigator of the Patrick and Catherine Weldon Donaghue Medical Research Foundation.

Correspondence and requests for materials should be addressed to S.M.S. (e-mail: stephen.strittmatter@yale.edu).

Torque-generating units of the flagellar motor of *Escherichia coli* have a high duty ratio

William S. Ryu*, Richard M. Berry† & Howard C. Berg*

* Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA and Rowland Institute for Science, 100 Edwin H. Land Boulevard, Cambridge, Massachusetts 02142, USA

† The Randall Centre, King's College London, London SE1 1UL, UK

Rotation of the bacterial flagellar motor is driven by an ensemble of torque-generating units containing the proteins MotA and MotB^{1–3}. Here, by inducing expression of MotA in *motA*[–] cells under conditions of low viscous load, we show that the limiting speed of the motor is independent of the number of units: at vanishing load, one unit turns the motor as rapidly as many. This result indicates that each unit may remain attached to the rotor for most of its mechanochemical cycle, that is, that it has a high duty ratio⁴. Thus, torque generators behave more like kinesin, the protein that moves vesicles along microtubules, than myosin, the protein that powers muscle. However, their translation rates, stepping frequencies and power outputs are much higher, being greater than 30 $\mu\text{m s}^{-1}$, 12 kHz and $1.5 \times 10^5 \text{ pN nm s}^{-1}$, respectively.

Bacterial flagellar motors are driven by a transmembrane ion flux (reviewed in refs 5, 6). Most studies of their physiology have been

done at low speeds, characteristic of high viscous loads, where translation rates of ions or movement of internal mechanical components are not limiting. To study their behaviour at higher speeds, we developed a new motor assay (Fig. 1a, inset). Instead of tethering a cell to a glass surface by a single flagellum and watching the cell body spin⁷, we fixed the cell body to the glass surface and attached a polystyrene bead to a stub of one of its flagellar filaments. By using beads of various sizes (0.3–1 μm diameter), we could vary the load by a factor of more than 10. The bead was followed in a weak optical trap by back focal plane interferometry⁸.

Figure 1a, b shows speed records for cells under high and low load, respectively. Upon induction of MotA expression, discrete increments in speed were observed. At high load (Fig. 1a), the increase in speed was linear with the torque-generator number, N , as observed previously with tethered cells^{1,2}. However, at lower load (Fig. 1b), the speed tended to saturate at high torque-generator numbers. The data for different bead sizes are summarized in Fig. 2a. In Fig. 3a we show torque–speed curves for motors with between one and five torque-generating units, constructed from the data of Fig. 2a. At low speeds, the torque produced by a motor with N generators is simply N times the torque produced by a motor with one generator. At high speeds, the torques decline, and the curves are consistent with a limiting speed at zero torque of about 300 Hz.

The behaviour of the motor can be understood in terms of the ‘duty ratio’ of the torque generators⁴, where the duty ratio, D , is defined as the fraction of time for which a generator is bound to the rotor. As the motor works against a viscous load, speed is proportional to the applied torque. At high loads, speed increases linearly with N whether D is small (close to 0) or large (close to 1). If D is small, the torque-generating units work independently of one another, because the probability that two might be attached at the same time is small; speed is proportional to the time-averaged torque. If D is large, each unit has time to reach thermodynamic equilibrium and exert the same torque; speed is proportional to the total torque. At vanishing load, speed increases linearly with N when

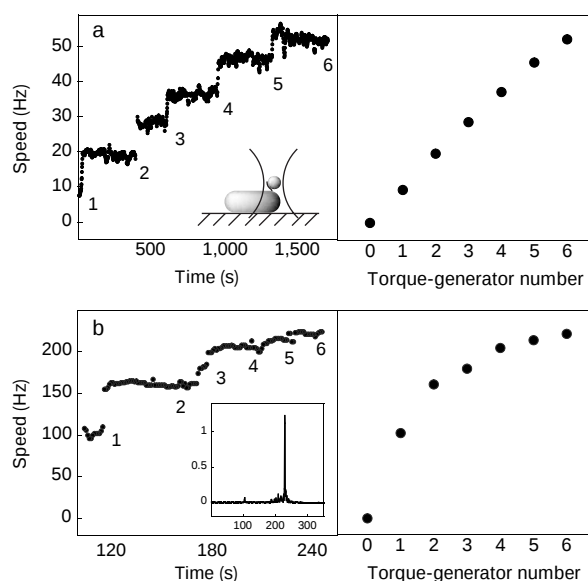


Figure 1 Measurements made at high and low load. **a**, Induction of torque in a motor driving a bead of 1.03 μm diameter (left). Speed is shown as a function of time, and the number of torque generators is indicated. Inset, a small bead attached to a segment of a flagellar filament and followed in a weak optical trap. Right, mean speed as a function of torque-generator number. Error bars are smaller than data markers. **b**, As in **a** for a different motor driving a bead of 0.36 μm diameter. Inset, the power spectrum of the signal from 80 revolutions of a 0.36- μm bead. The abscissa is frequency (Hz), and the ordinate is power (arbitrary units).

D is small but not when D is large. If D is small, the torque-generating units work independently of one another, as before, and again speed is proportional to the time-averaged torque. If D is large, on the other hand, any unit bound to the rotor limits its speed. Assuming that the time required for a unit to complete a mechanochemical cycle is independent of the torque exerted by other units, the speed is determined by the displacement per cycle divided by the cycle time. So our data are consistent with a model in which the duty ratio is close to 1.

Our results can be reproduced by a model (Box 1) in which $D = 1$ and the rate constants for forward transitions are load independent, as shown in Figs 2b and 3b. If an unbound state is introduced ($D < 1$) or if all forward rate constants are load-dependent, the zero-load speeds vary with N (Box Fig. b) and the concave-down shape of the torque-speed curves^{9,10} cannot be reproduced. Our data do not rule out zero-load speeds that vary with N , as it was not possible to measure these speeds directly—beads had to be at least $0.30\ \mu\text{m}$ in diameter to detect rotation. By extrapolating the curves for $N = 3, 4$ and 5 unit motors to zero torque (using the fastest three data points of each curve, Fig. 3a), we estimate the zero-load speeds to be $306 \pm 76\ \text{Hz}$, $299 \pm 47\ \text{Hz}$ and $305 \pm 35\ \text{Hz}$, respectively. Combining these values we obtain $303 \pm 32\ \text{Hz}$. The ratio of the lowest to the highest zero-load speeds within the error limits is ~ 0.8 . When this ratio is compared to the corresponding ratios of zero-load speeds obtained from simulations at various duty ratios (Box Fig. b), we find $D > 0.6$.

We conclude that the torque-generating units remain attached to the rotor most of the time; in other words, that their duty ratio is greater than 0.6 and more likely to be closer to 1. The speed might saturate at low loads at a lower duty ratio if the generators work cooperatively at low loads and bind synchronously; however, there

is at present no evidence for cooperative interactions between units. Alternatively, a unit that appears to have a duty ratio close to 1 might be composed of more than one interacting subunit with a lower duty ratio, as in the dimeric kinesin molecule. If so, our analysis applies to the composite structure, not the individual subunits.

It has been found from fluctuation analysis of tethered cells rotating at low speed that the total number of steps taken by the rotor per revolution is proportional to torque-generator number, N (ref. 3). This implies that the rotor step size decreases with N . This could be understood if, on stepping along the periphery of the rotor a fixed distance d , the N th torque-generating unit had to work with the other $N - 1$, suggesting a high duty ratio. Our work supports this argument and shows that the duty ratio remains high for speeds up to a few hundred hertz. Thus, torque generators behave more like kinesin than myosin. Single kinesin molecules take hundreds of steps along a microtubule without falling off¹¹, whereas myosin spends most of its mechanochemical cycle detached from the actin filament, binding only briefly to undergo a force-producing step¹². In both cases, the function of the motor molecule dictates the duty ratio. When vesicles are moved along microtubules, procession is more important than speed, so kinesin has a high duty ratio. When muscles contract, speed is more important than procession, so many myosins work together, each with a low duty ratio¹³. Evidently, for the flagellar motor, the ability to drive sizeable viscous loads at high speeds requires that the torque generators work with a high duty ratio. For *E. coli*, this probably relates to the forces required to form or disrupt flagellar bundles, processes essential for chemotaxis¹⁴. However, a high duty ratio requires that each generator completes a fixed number of torque-generating cycles each revolution, regardless of speed, so high speeds require high

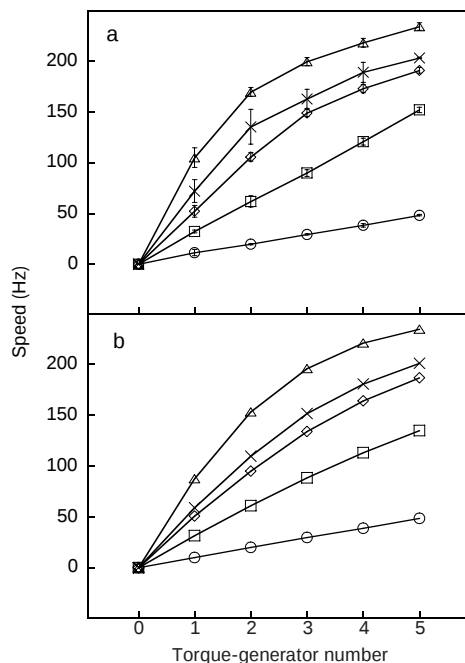


Figure 2 Speeds, measured and simulated. **a**, Speed as a function of torque-generator number for different bead sizes. Reading from top to bottom, the diameter of the beads in micrometres (and the numbers of motors assayed at each bead size) were 0.30 (12), 0.356 (15), 0.41 (8), 0.535 (11) and 1.03 (3), respectively. The data points and error bars are the means and standard deviations over the cell ensemble, with the mean speeds for each torque-generator number weighted equally. **b**, A Monte-Carlo simulation of the behaviour of a motor shown as a function of torque-generator number, N . Reading from top to bottom, the viscous drag coefficients were 2.7, 4.3, 5.1, 8.5 and $28.0\ \text{pN nm}$ per revolution per s.

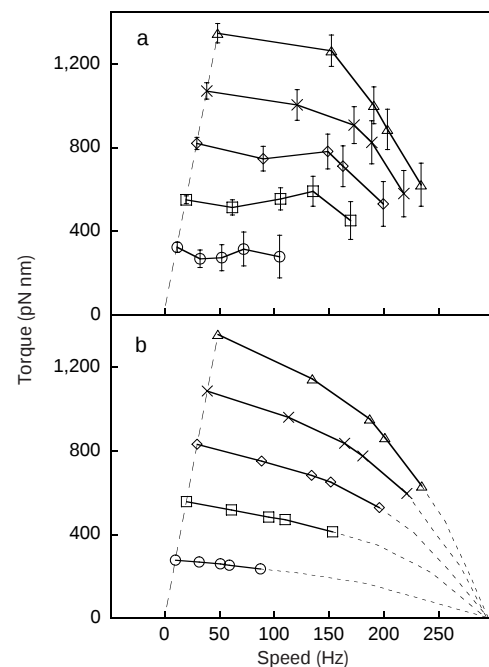


Figure 3 Torques, measured and simulated. **a**, Torque-speed curves for motors with 1 (bottom) to 5 (top) torque-generating units. The curves were derived from the data of Fig. 2a, as described in Methods. The points for a given bead size fall on a line—one such line is shown—whose slope is equal to the viscous drag coefficient. Reading from left to right, the viscous loads were 27.8, 8.5, 5.3, 4.4 and $2.7\ \text{pN nm}$ per revolution per s. **b**, A similar plot derived from the simulation of Fig. 2b, including points computed at lower viscous loads (dashed lines).

Box 1

Flagellar motor model

A model was developed to reproduce the torque–speed curves, which are concave downwards^{9,10}. Torque generators work independently with a three-state cycle, $E_p \leftrightarrow A_p \leftrightarrow B_p \leftrightarrow E_{p+1}$, where p identifies the binding site²⁸. Each cycle carries one proton across the membrane and moves the equilibrium position of the rotor by $\varphi = 2\pi/n_r$, where n_r is the number of binding sites. Proton exchange with the external and internal media is represented by transitions $E_p \leftrightarrow A_p$ and $B_p \leftrightarrow E_{p+1}$, respectively, and changes that occur while protons are in the motor are represented by the transition $A_p \leftrightarrow B_p$. For each state ($i = E_p, A_p, B_p$) the energy $U_i(\theta)$ represents the elastic, electrostatic, and/or hydrophobic energies involved in generator/rotor interaction. Each proton provides the energy $U_H = qV_m$, where q is the proton charge and V_m the protonmotive force, 150 mV. Rate constants for forwards and backwards transitions are given by $k_{ij} = k^0 \exp(\lambda \Delta G_{ij}/k_b T)$ and $k_{ji} = k^0 \exp[(\lambda - 1) \Delta G_{ij}/k_b T]$, respectively. $\Delta G_{ij} = U_i(\theta) - U_j(\theta) + \alpha_{ij} U_H$, where α_{ij} is the fraction of U_H dissipated in the transition, λ varies from 0 to 1 and specifies how much changes in ΔG affect forward as opposed to backward rate constants, k^0 controls the intrinsic rate of the transition, k_b is Boltzmann's constant and T is the absolute temperature. Each unit exerts a torque $T = -dU_i/d\theta$, and the total motor torque is the sum over all units. Transitions between states and the stochastic equations for rotation of the rotor and bead attached via an elastic link (proximal hook stiffness, $k_\theta = 400$ pN nm per radian)^{29,30} are modelled by Monte-Carlo simulation.

The following model parameters were necessary to reproduce the measured torque–speed curves^{9,10}. A powerstroke mechanism was chosen, with 90% of U_H and 90% of φ placed in the transition $A \leftrightarrow B$ ²⁸. The rest of U_H and φ was split evenly between the other transitions. λ

were all zero and k^0 were 1.23×10^5 s⁻¹. Energy profiles, $U_i(\theta)$, for all three states were those of a spring that becomes less stiff at large extensions. That is, for positive extensions, ψ , the torque was given by

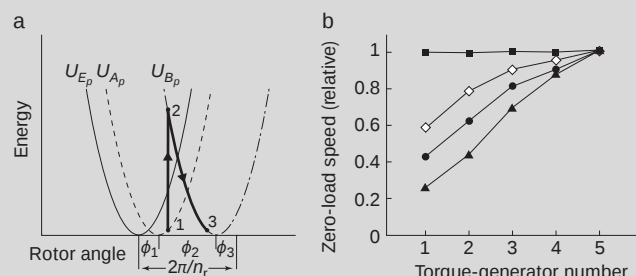
$$-\frac{dU}{d\theta} = \begin{cases} -\kappa_1 \psi & : \psi < a \\ -[\kappa_1 a + \kappa_2(\psi - a)] & : \psi > a \end{cases}$$

where $\kappa_2 < \kappa_1$. For negative extensions (extensions that tend to slow the motor) the torque had the opposite sign. The predictions of the model were not very sensitive to the values of κ_1 , κ_2 and a .

An example of a transition is shown in the energy diagram, a. A torque-generating unit in state A_p is in mechanical equilibrium with the rotor at 1. A forward transition to state B_p moves the equilibrium position through an angle ϕ_2 , placing the unit out of mechanical equilibrium at 2. Torque $T = -dU_B(\theta)/d\theta$ is produced in the forward direction, until equilibrium is reached at 3. An earlier transition from state E_p to A_p moved the equilibrium position an angle ϕ_1 . The next transition to state E_{p+1} will move the equilibrium position through an angle ϕ_3 to the next binding site $p + 1$, completing the cycle.

For simulations with $D < 1$, an unbound state that produced zero torque was introduced by setting $U_E(\theta)$ to be constant, and the rates into and out of this state were controlled by choosing the value of U_E . When the torque-generating units are unbound, the rotor can diffuse freely and respond to external torques, so the rotational drag coefficient of the rotor becomes important. When torque is exerted by the motor on the external load (bead), the tether winds up. When the units are unbound, the tether unwinds and moves the rotor backwards, implying a loss of efficiency. This happens because the viscous drag of the bead is much larger than that of the rotor. The rotational drag of the rotor was estimated to be $f_r = 0.25$ pN nm per (rev per s), modelled as a 50 nm diameter sphere in a lipid of viscosity 100 cp, yielding a relaxation time $\tau_r = f_r/k_\theta \approx 1$ ms.

Figure b shows the dependence of speed on the number of torque-generating units at zero load for different simulation parameters. The speeds have been normalized individually for each curve by the speed at $N = 5$; therefore, comparisons of speed cannot be made between curves. For units with $D = 1$ (squares), the zero-load speed is independent of N . For units with $D < 1$, the zero-load speed depends on N (circles, $D = 0.6$; triangles, $D = 0.23$). For units with load-dependent forward rate constants (diamonds, $D = 1$, $\lambda = 0.5$ and U_H and φ split equally between all three transitions), the zero-load speed also depends on N , irrespective of the duty ratio.



turnover rates. For example, a motor with $N = 5$ can spin a 0.30- μ m bead at 240 Hz (Fig. 2a). If each unit steps 50 times per revolution—this is a lower bound³—the cycle frequency is at least 12 kHz. As the rotor is about 50 nm in diameter¹⁵, the speed at the periphery exceeds 35 μ m s⁻¹. Each unit can produce more than 5 pN of force and more than 1.75×10^3 pN nm s⁻¹ of power. The Na⁺-driven flagellar motor in *Vibrio alginolyticus* rotates even more rapidly (1,000 Hz), indicating that higher turnover rates and power outputs may be possible¹⁶. In comparison, a single kinesin molecule at saturating ATP concentration produces over 70 times less power, moving at about 0.5 μ m s⁻¹ while exerting a force of 5 pN (ref. 17). A single myosin molecule produces over 400 times less power, stepping 20 nm at a turnover rate of 10 s⁻¹, and exerting an average force of 2 pN (ref. 18). F₁-ATPase produces over 450 times less power, rotating at about 2 Hz against a load torque of 30 pN nm (ref. 19). The reason for the higher numbers in *E. coli* might be because the flagellar motor is driven by translocation of ions rather than binding, hydrolysis and release of nucleotides. □

Methods

Cells and beads

We used strain HCB1271. It was constructed from strain MS5037 (ref. 20), containing *motA448* (ref. 21). MS5037 was made pil-minus by P1 transduction of pilA'-kan from

strain HCB758 (refs 22, 23) and filament-minus by transduction of *fliC*::Tn10 from strain C600 *hsm hsr fliC*::Tn10 (ref. 24) by selection with the appropriate antibodies. We introduced *MotA* resurfacing capability by transformation with pDFB36, carrying wild-type *motA* under an inducible promoter². The *fliC*st (sticky filament) allele²⁵ was cloned into a plasmid compatible with pDFB36. Wild-type *MotA* was induced on addition of isopropyl β -D-thiogalactopyranoside (IPTG), and the sticky filaments readily adsorbed polystyrene beads. Reconstituted motors spun alternately clockwise and counterclockwise, but only the latter intervals were used here. Cells were prepared as described⁹, but we sheared more completely by 60 passages through the shearing apparatus and resuspended at half the original volume in motility medium (10 mM potassium phosphate, pH 7.0, 0.1 mM EDTA). The cells were allowed to settle on a glass coverslip that had been treated with 4-aminobutyltrimethoxy silane⁶. After 15 min, unattached cells were washed away, and a suspension of polystyrene beads was added (0.083% w/w in motility medium; 1.03, 0.525, 0.356 or 0.30 μ m diameter, from Polysciences, or 0.41 μ m diameter from Bangs Laboratories). After 5 min, unattached beads were washed away and the tethering medium was supplemented with tryptone broth (1% tryptone, Difco; 0.5% NaCl) at a concentration up to 20% v/v and with IPTG (final concentration 0.5 μ M to 2.5 mM).

Data acquisition and analysis

The optical trap is described elsewhere²⁶. We used two functionally identical instruments, one with a 980-nm MOPA laser²⁶ and the other with a 830-nm diode laser (Melles Griot). The power level at the back focal plane of the objective was 20 mW (980 nm, a wavelength at which the objective had a low transmittance) and <10 mW (830 nm). Trapping forces were too small to perturb the motion. Signals were sampled at 0.5–1 kHz and anti-alias filtered at 250 Hz. Power spectra were computed for successive 1-s blocks of data, and the speed–time plots were smoothed with a median filter of rank 4, which preserved discontinuities in the data. Decisions on the correspondence between speed and torque-

generator number were made at speeds below 160 Hz, where speed increased linearly with generator number.

Construction of torque-speed plots

We calculated torque, T , from estimates of drag coefficients, f , and measurements of speed. Rotational drag caused by an eccentrically rotating bead (of radius r_b rotating with an eccentricity r_e) attached to a flagellar stub (modelled as a thin prolate ellipsoid with a major radius a and minor radius b (10 nm) moving edge-on at a radius r_f) was estimated to be $8\pi\eta(r_b^3 + r_e^3s) + 6\pi\eta r_e^2r_b$, where s is the structural factor $a/[\ln(2a/b) + 0.5]$ (ref. 27). From digitized video images, the average eccentricity of rotation was $0.15 \pm 0.5 \mu\text{m}$. The rotational radius of the flagellar stub, r_b , which is proportional to stub length, was set to $a/2$. The average flagellar stub length, $2a$, was unknown and was left as a free parameter for subsequent fitting. The drag for each bead size was adjusted so that the data points for $N = 5$ generators coincided with a wild-type ($N = 8$) torque-speed curve^{9,10}, scaled to 5/8 the stall torque, but with the same knee and zero-torque speeds. The resulting torque-speed curves are shown in Fig. 3a. The fitted flagellar stub lengths were 1.22, 1.07, 0.95, 0.91 and $0.75 \mu\text{m}$, for bead sizes 1.0, 0.54, 0.41, 0.36 and $0.30 \mu\text{m}$, respectively.

Received 6 July; accepted 22 November 1999.

- Block, S. M. & Berg, H. C. Successive incorporation of force-generating units in the bacterial rotary motor. *Nature* **309**, 470–473 (1984).
- Blair, D. F. & Berg, H. C. Restoration of torque in defective flagellar motors. *Science* **242**, 1678–1681 (1988).
- Samuel, A. D. T. & Berg, H. C. Torque-generating units of the bacterial flagellar motor step independently. *Biophys. J.* **71**, 918–923 (1996).
- Leibler, S. & Huse, D. A. Porters versus rowers: a unified stochastic model of motor proteins. *J. Cell Biol.* **121**, 1357–1368 (1993).
- Macnab, R. M. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* Vol. 1 (eds Neidhardt, F. C. et al.) 123–145 (American Society for Microbiology, Washington, 1996).
- Berry, R. M. & Armitage, J. P. The bacterial flagellar motor. *Adv. Microb. Physiol.* **41**, 291–337 (1999).
- Silverman, M. & Simon, M. Flagellar rotation and the mechanism of bacterial motility. *Nature* **249**, 73–74 (1974).
- Gittes, F. & Schmidt, C. F. Interference model for back-focal-plane displacement detection in optical tweezers. *Opt. Lett.* **23**, 7–9 (1998).
- Berg, H. C. & Turner, L. Torque generated by the flagellar motor of *Escherichia coli*. *Biophys. J.* **65**, 2201–2216 (1993).
- Chen, X. & Berg, H. C. Torque-speed relationship of the flagellar rotary motor of *Escherichia coli*. *Biophys. J.* (in the press).
- Block, S. M., Goldstein, L. S. B. & Schnapp, B. J. Bead movement by single kinesin molecules studied with optical tweezers. *Nature* **348**, 348–352 (1990).
- Uyeda, T., Abramson, P. D. & Spudis, J. A. The neck region of the myosin motor domain acts as a lever arm to generate movement. *Proc. Natl Acad. Sci. USA* **93**, 4459–4464 (1996).
- Howard, J. Molecular motors: structural adaptations to cellular functions. *Nature* **389**, 561–567 (1997).
- Macnab, R. M. Bacterial flagella rotating in bundles: a study in helical geometry. *Proc. Natl Acad. Sci. USA* **74**, 221–225 (1997).
- DeRosier, D. J. The turn of the screw: the bacterial flagellar motor. *Cell* **93**, 17–20 (1998).
- Muramoto, K. et al. High-speed rotation and speed stability of the sodium-driven flagellar motor in *Vibrio alginolyticus*. *J. Mol. Biol.* **251**, 50–58 (1995).
- Visscher, K., Schnitzer, M. J. & Block, S. M. Single kinesin molecules studied with a molecular force clamp. *Nature* **400**, 184–189 (1999).
- Toyoshima, Y. Y. et al. Myosin subfragment-1 is sufficient to move actin filaments *in vitro*. *Nature* **328**, 536–539 (1987).
- Yasuda, R., Noji, H., Kinosita, K. J. & Yoshida, M. F₁-ATPase is a highly efficient molecular motor that rotates with discrete 120° steps. *Cell* **93**, 1117–1124 (1998).
- Silverman, M. & Simon, M. Operon controlling motility and chemotaxis in *E. coli*. *Nature* **264**, 577–580 (1976).
- Armstrong, J. B. & Adler, J. Genetics of motility in *Escherichia coli*: complementation of paralyzed mutants. *Genetics* **56**, 363–373 (1967).
- Fung, D. C. Y. K. *Powering the Escherichia coli Flagellar Motor with an External Voltage Source*. Thesis, Harvard Univ. (1994).
- Maurer, L. & Orndorff, P. E. Identification and characterization of genes determining receptor binding and pilus length of *Escherichia coli* type 1 pili. *J. Bacteriol.* **169**, 640–645 (1987).
- Kuwajima, G. Flagellin domain that affects H antigenicity of *Escherichia coli* K-12. *J. Bacteriol.* **170**, 485–488 (1988).
- Scharf, B. E., Fahrner, K. A., Turner, L. & Berg, H. C. Control of direction of flagellar rotation in bacterial chemotaxis. *Proc. Natl Acad. Sci. USA* **95**, 201–206 (1998).
- Berry, R. M. & Berg, H. C. Absence of a barrier to backwards rotation of the bacterial flagellar motor demonstrated with optical tweezers. *Proc. Natl Acad. Sci. USA* **94**, 14433–14437 (1997).
- Happel, J. & Brenner, H. *Low Reynolds Number Hydrodynamics* (Kluwer Academic, Dordrecht, 1991).
- Berry, R. M. & Berg, H. C. Torque generated by the flagellar motor of *Escherichia coli* while driven backward. *Biophys. J.* **76**, 580–587 (1999).
- Block, S. M., Blair, D. F. & Berg, H. C. Compliance of bacterial flagella measured with optical tweezers. *Nature* **338**, 514–517 (1989).
- Block, S. M., Blair, D. F. & Berg, H. C. Compliance of bacterial polyhooks measured with optical tweezers. *Cytometry* **12**, 492–496 (1991).

Acknowledgements

We thank K. Fahrner and A. Samuel for plasmid and strain constructions, and K. Fahrner, A. Samuel and L. Turner for comments on the manuscript. This work was supported by the Wellcome Trust, the Rowland Institute for Science and by the US NIH.

Correspondence and requests for materials should be addressed to H.C.B.

The DExH protein NPH-II is a processive and directional motor for unwinding RNA

Eckhard Jankowsky*, Christian H. Gross†, Stewart Shuman† & Anna Marie Pyle*‡

* The Department of Biochemistry and Molecular Biophysics, Columbia University, 630 W. 168th St, New York, New York 10032, USA

† The Program in Molecular Biology, Sloan-Kettering Institute,

‡ The Howard Hughes Medical Institute 1275 York Ave., New York, New York 10021, USA

All aspects of cellular RNA metabolism and processing involve DExH/D proteins, which are a family of enzymes that unwind or manipulate RNA in an ATP-dependent fashion¹. DExH/D proteins are also essential for the replication of many viruses, and therefore provide targets for the development of therapeutics². All DExH/D proteins characterized to date hydrolyse nucleoside triphosphates and, in most cases, this activity is stimulated by the addition of RNA or DNA¹. Several members of the family unwind RNA duplexes in an NTP-dependent fashion *in vitro*^{1,3}; therefore it has been proposed that DExH/D proteins couple NTP hydrolysis to RNA conformational change in complex macromolecular assemblies⁴. Despite the central role of DExH/D proteins, their mechanism of RNA helicase activity remains unknown. Here we show that the DExH protein NPH-II unwinds RNA duplexes in a processive, unidirectional fashion with a step size of roughly one-half helix turn. We show that there is a quantitative connection between ATP utilization and helicase processivity, thereby providing direct evidence that DExH/D proteins can function as molecular motors on RNA.

Establishing a basic functional model for RNA helicase behaviour involved answering three fundamental questions: first, does the helicase unwind RNA in distinct, sequential steps? Second, does the enzyme dissociate from the RNA after each unwinding step or does it stay bound; that is, is the helicase processive? Third, does the helicase move from one defined point on its helical substrate to another defined point; is unwinding unidirectional?

We subjected the DExH helicase NPH-II from vaccinia virus⁵ to quantitative kinetic analysis of RNA helicase activity using a collection of duplex RNA substrates that varied in composition and length. Kinetics of unwinding were monitored using gel-shift electrophoresis and fluorescence energy transfer methodologies. The RNA substrates contained a single-strand 3' overhang, which is required for NPH-II unwinding activity⁶. The single-strand 3' overhang was identical in length and sequence for every substrate (Fig. 1). Unwinding reactions were performed under single-turnover conditions with respect to the RNA substrate. A large excess of nonspecific 'trap RNA' was added to prevent helicase from re-associating with duplex once it falls off during the course of reaction.

In the absence of trap RNA, the unwinding rate and reaction amplitude (defined as the final fraction of unwound RNA) were both insensitive to duplex length (Fig. 1). Each reaction was first-order with a rate constant of $3.5 \pm 0.2 \text{ min}^{-1}$ (Fig. 1). In the presence of trap RNA, the rates remained independent of duplex length; however, the reaction amplitude decreased with increasing duplex length (Fig. 1). These observations provide three mechanistic insights. First, the fact that long duplexes can be unwound at all in the presence of trap RNA provides strong qualitative evidence that the helicase is a processive enzyme, consistent with previous multiple-turnover studies⁷. Second, the dependence of amplitude on duplex length indicates that more protein dissociates from an

RNA substrate when the duplex is longer, indicating that more unwinding steps may be required for a longer duplex, and there is a specific degree of processivity for the NPH-II enzyme (Fig. 1). Last, the similar rates observed for the unwinding of different duplex lengths indicate that individual unwinding steps do not limit the reaction rate under these conditions. Rate-limiting events during translocation would result in a length-dependent lag phase in each time course⁸.

To evaluate helicase step size and to provide additional proof of processivity, it was necessary to monitor individual unwinding steps. We therefore sought conditions under which helicase translocation becomes rate limiting. Variation in temperature and pH did not cause the rates to become increasingly sensitive to duplex length (E.J. *et al.*, unpublished data); however, marked effects were observed when Mg^{2+} , which is the cofactor for ATP hydrolysis, was replaced by Co^{2+} . The unwinding reactions displayed pronounced lag phases that increased incrementally with the duplex length (Fig. 2). The data were fit numerically to a kinetic model that relates the translocation rate constant to step size and duplex length (Methods). From this analysis, the unwinding step size of NPH-II was estimated to be six base pairs (bp) and the unwinding rate constant (for each translocative step in Co^{2+} at an ATP concentration of 3.5 mM), k_U , was $13.4 \pm 0.8 \text{ min}^{-1}$. Using the step size of 6 bp, it is possible to calculate a lower limit for the translocation rate constant in Mg^{2+} . In these conditions, k_U was greater than 350 min^{-1} ,

indicating that translocation in Mg^{2+} is at least 25-fold faster than in Co^{2+} . The slower unwinding rate in Co^{2+} is not attributable to inhibition of ATPase activity, as kinetic parameters for hydrolysis of ATP are similar in Co^{2+} and Mg^{2+} (E.J. *et al.*, unpublished data).

The existence of a defined step size indicates regularity in the translocation process; that is, unwinding is not random with respect to individual translocation rates and the number of bp displaced in one unwinding step. The step size obtained herein corresponds to about half of a helical turn and is very similar to the unwinding step size of the UvrD DNA helicase of 4–5 bp (ref. 8). It is important to note that the unwinding step size is an average value that may not necessarily reflect actual microscopic translocation steps of a helicase⁸.

Although the above results suggest that RNA unwinding occurs in consecutive steps, the assays that we used are designed to monitor only the very last event in strand displacement (one does not see a signal until the duplex is completely unwound). Therefore, it is impossible to distinguish between directional translocation and translocation that occurs through random facilitated diffusion^{9–11}. To determine which of these mechanisms applies, we needed to detect an unwinding intermediate. This was accomplished by using multipiece substrates (MPS), consisting of a continuous bottom strand that is annealed to two adjacent pieces of top-strand RNA (Fig. 3a). The top-strand pieces were generated by site-specific cleavage of a single continuous piece of top-strand RNA using synthetic DNAzyme endonucleases¹². The two resultant pieces had different lengths and could therefore be distinguished by their respective electrophoretic mobilities (Fig. 3b, c). To prevent the results from being biased by the relative length of the top-strand pieces, we designed two different sets of MPS (Fig. 3a). All MPS unwinding reactions were carried out under single-turnover conditions, with respect to the RNA substrate, and in the presence of trap RNA.

NPH-II unwound both sets of MPS (Fig. 3b), indicating that the helicase can proceed through nicks in the top strand of the substrate. In the presence of Mg^{2+} , all pieces were displaced with apparent first-order kinetics with similar rate constants, $k = 3.3 \pm 0.5 \text{ min}^{-1}$ (Fig. 3b). For both sets of MPS, the piece that was bound farthest from the overhang was displaced with a lower amplitude than the

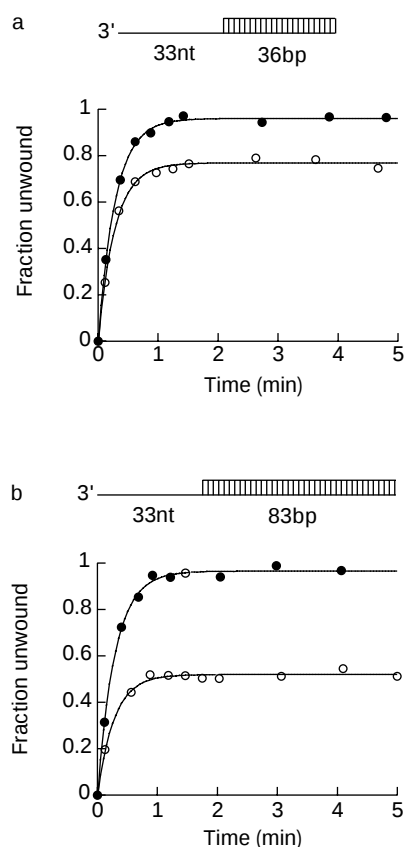


Figure 1 Kinetics of duplex unwinding. Reactions in the presence (open circles) and absence (filled circles) of trap RNA with a 36-bp substrate (**a**) and an 83-bp substrate (**b**). The 83-bp substrate is an extension of the 36-bp substrate. The fraction of unwound substrate (indicated on the y axis) was fitted to the integrated first-order rate law using Kaleidagraph software (Abelbeck). Rate constants (k) and reaction amplitudes (A) in the absence of trap RNA: $k = 3.5 \pm 0.1 \text{ min}^{-1}$, $A = 0.96$, 36-bp substrate; $k = 3.4 \pm 0.1 \text{ min}^{-1}$, $A = 0.97$, 83-bp substrate. In the presence of trap RNA: $k = 3.6 \pm 0.2 \text{ min}^{-1}$, $A = 0.77$, 36-bp substrate; $k = 3.6 \pm 0.3 \text{ min}^{-1}$, $A = 0.52$, 83-bp substrate. Variances are standard deviations from the fit.

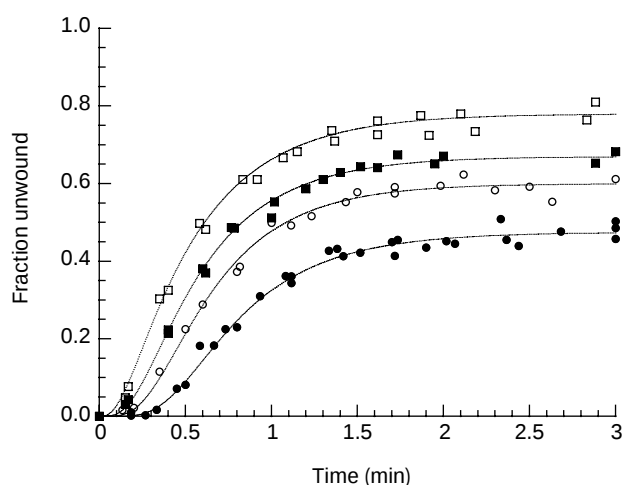


Figure 2 Estimation of the unwinding step size. Time courses of substrates containing duplex regions of 12 (open squares), 18 (filled squares), 24 (open circles) and 36 (filled circles) bp unwound by NPH-II in the presence of 4 mM $CoCl_2$. Substrates contain a 35-U single-strand overhang 3' to the duplex region. Step size and the unwinding (translocation) rate constant were calculated from at least two independent time courses for each duplex, which are overlaid on the plots. Solid lines are best fits to the kinetic model, resulting in an unwinding step size of 6 bp and an average unwinding rate constant of $k_U = 13.4 \pm 0.8 \text{ min}^{-1}$.

piece located closer to the single-strand 3' overhang (Fig. 3b). This implies that fewer unwinding steps were required for the displacement of the piece closer to the overhang. The fact that these effects depend exclusively on the location of the top strands relative to the single-strand 3' overhang (rather than their length) strongly suggests that the strand displacement progresses in a distinct 3'-to-5' direction with respect to the bottom strand.

The similar unwinding rates for both pieces of top-strand RNA indicate that neither a late step in the reaction, such as the final strand separation, nor translocation itself, limits the rate of the overall process in Mg^{2+} . A late rate-limiting step would result in a slower rate for unwinding of the piece farthest from the overhang because two strand separations are required for unwinding of the second piece. A rate-limiting translocation would cause a lag phase in the time courses⁸, as observed in the presence of Co^{2+} (Fig. 2). Thus, in the presence of Mg^{2+} , an initiation step at the very beginning of the unwinding process limits the rate of the overall reaction.

In contrast, reactions conducted in the presence of Co^{2+} displayed a pronounced lag phase attributable to rate-limiting translocation steps (Fig. 3c). Notably, the displacement of the two pieces occurred with different lag phases. The lag was largest for the piece farthest from the overhang, reflecting an increase in the number of unwinding steps that are required as the helicase moves away from the single-strand 3' overhang. The effects were exclusively dependent

on the location of the pieces relative to the overhang and not on their length, indicating a distinct 3'-to-5' directionality in the progression of unwinding with respect to the bottom strand.

Having established that the helicase unwinds RNA substrates in a processive and directional fashion with a distinct unwinding step size, it was of interest to correlate this process with the role of ATP. The effect of ATP on enzyme processivity (P) was of particular interest (Fig. 4). Processivity reflects the probability that the helicase will perform the next unwinding step rather than dissociate from the substrate¹³. Quantitation of processivity was carried out by determining the reaction amplitude in the presence of trap RNA, such that all dissociation of protein from the RNA substrate during unwinding was effectively irreversible. Processivity decreased with decreasing ATP concentration (Fig. 4a), which indicates that the helicase dissociates more readily from RNA in non-ATP-bound states than in the ATP-bound state, consistent with previous studies of RNA binding by NPH-II (ref. 5). In addition, the processivity at ATP saturation approaches $P = 1$, regardless of whether Co^{2+} or Mg^{2+} are the cofactors for ATP hydrolysis (Fig. 4a). This indicates that NPH-II in the ATP-bound state does not significantly dissociate from the RNA substrate, and that dissociation of protein from RNA occurs predominantly in non-ATP-bound states. This is highly instructive because it shows that ATP binding, and not just the consumption of ATP through hydrolysis, is of critical importance in the helicase mechanism.

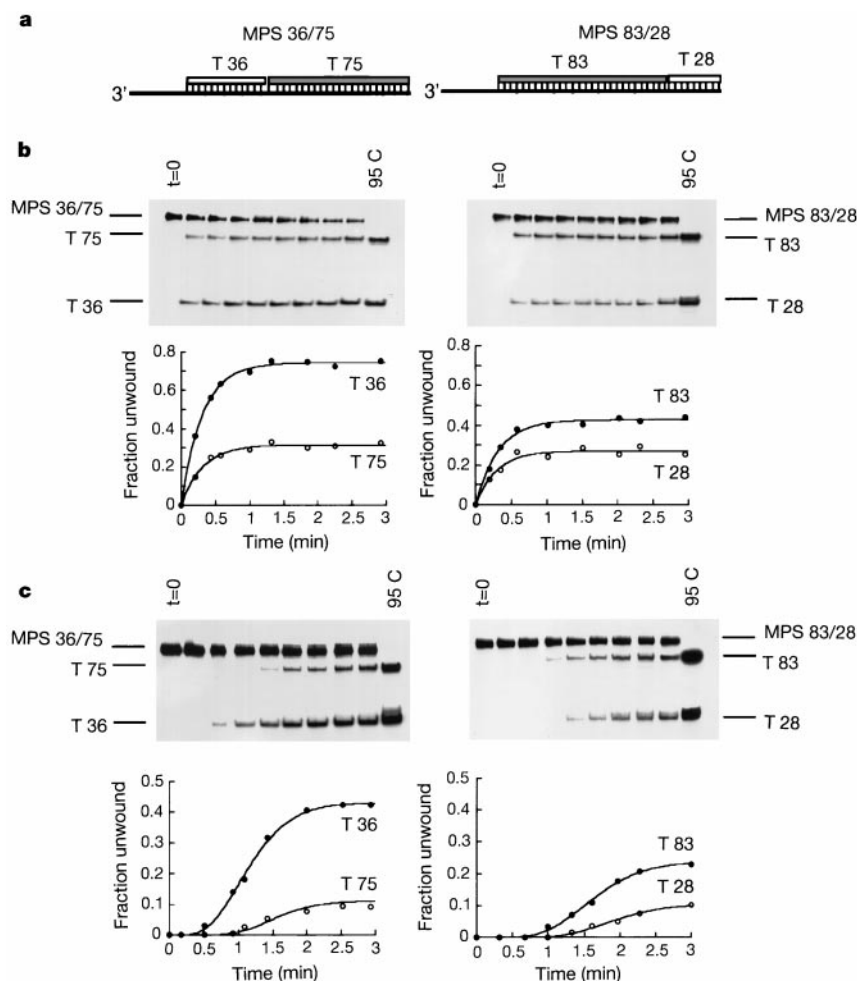


Figure 3 Probing helicase directionality with multipiece substrates (MPS). **a**, Design of the MPS. **b**, Unwinding of MPS in 4 mM $MgCl_2$ and trap RNA. Unwinding rate constants were similar for all pieces ($k = 3.3 \pm 0.5 \text{ min}^{-1}$); reaction amplitudes (A) were MPS 36/75, T 36, $A = 0.74$; T 75, $A = 0.32$; MPS 83/28, T 83, $A = 0.43$; T 28, $A = 0.27$.

c, Unwinding of multipiece substrates in 4 mM $CoCl_2$ and trap RNA. Kinetic data were analysed as described in Fig. 2, which resulted in average values of $k_0 = 12.4 \pm 2.1 \text{ min}^{-1}$ and a step size of 6 bp, consistent with the values obtained for oligomers shown in Fig. 2.

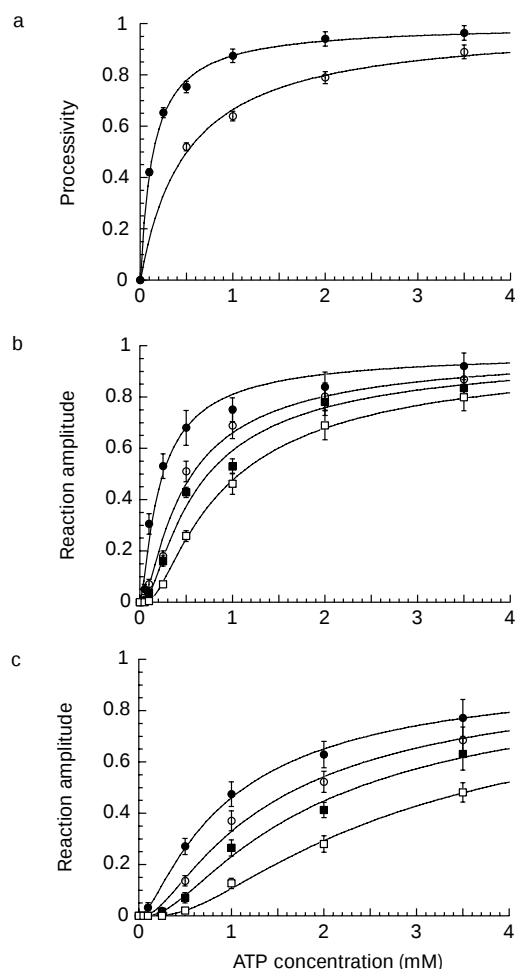


Figure 4 Processivity of NPH-II. **a**, ATP dependence of processivity in 4 mM MgCl₂ (filled circles) and 4 mM CoCl₂ (open circles). The solid line is a hyperbolic fit: $P = P_{\infty}[ATP]/(Z + [ATP])$, where $P_{\infty} = k_{U(S)}/(k_{U(S)} + k_F)$ reflects the processivity at ATP saturation; Z is the ATP concentration where $P = P_{\infty}/2$. In 4 mM CoCl₂, $P_{\infty} = 1.00 \pm 0.03$ and $Z = 0.50 \pm 0.06$ mM. In 4 mM MgCl₂, $P_{\infty} = 0.99 \pm 0.01$ and $Z = 0.14 \pm 0.01$ mM. Processivities were determined from the reaction amplitudes of four duplexes (12, 18, 24 and 36 bp, Fig. 2) at the ATP concentrations indicated. Amplitudes were obtained from at least three independent experiments. Processivity was calculated by fitting plots of amplitude at constant ATP concentration versus duplex length to equation (2) using a step size of $m = 6$ (Fig. 2). **b, c**, Dependence of reaction amplitude on ATP concentration for substrates with duplex regions of 12 (filled circles), 18 (open circles), 24 (filled squares) and 36 (open squares) bp in 4 mM MgCl₂ (**b**) and 4 mM CoCl₂ (**c**). Each reaction amplitude was determined in triplicate and the deviation from these independent measurements is indicated by the error bars. The solid lines are fits to equation 1, where K'_d was assumed to equal K_m from multiple turnover ATPase measurements (1.1 mM in Mg²⁺ and 3.5 mM in Co²⁺^{5,7}), the step size was 6 bp (Fig. 2), and $k_F/k_{U(S)}$ was allowed to float. In both Mg²⁺ and Co²⁺ $k_F/k_{U(S)} = 0.12$.

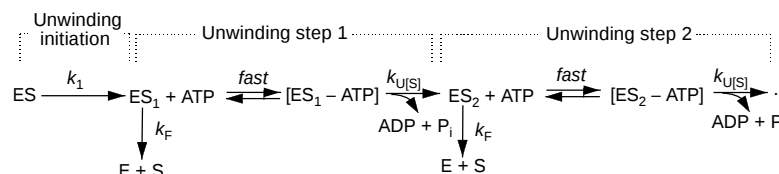


Figure 5 The kinetic scheme for RNA helicase activity by NPH-II. Unwinding initiation and two unwinding/translocation steps are depicted. ES₁ and ES₂ represent the enzyme-substrate complex in non-ATP-bound states, whereas [ES₁-ATP] and [ES₂-ATP] represent ATP-bound states. Note that 'ATP-bound' indicates that ATP is bound but not hydrolysed, whereas 'non-ATP-bound' indicates that no nucleotide is bound or that ADP and/or inorganic phosphate are bound. The equilibrium between ATP-bound and non-ATP-

Although RNA unwinding activity by NPH-II is dependent on ATP hydrolysis⁵, one must establish a quantitative link between ATP utilization and a specific amount of unwinding to conclude that DExH/D proteins behave as true molecular motors¹⁴. Plots of reaction amplitude versus ATP concentration resulted in sigmoidal curves (Fig. 4b, c) that are affected by duplex length: the longer the duplex, the more pronounced the sigmoidal shape of the time course. These results, together with the observation of a discrete step size and the fact that NPH-II does not dissociate in the ATP-bound state, are consistent with the kinetic scheme shown in Fig. 5. We used this scheme to derive an explicit equation that describes the plots of reaction amplitudes versus ATP concentration (Fig. 4b, c)

$$A = (1 - x) \cdot \left(1 + \frac{k_F}{k_{U(S)}} \cdot \frac{K'_d}{[ATP]} \right)^{-L/m} \quad (1)$$

The fit of this expression to the data establishes a direct relationship between processivity and ATP concentration, indicating that translocation is coupled to the binding of ATP. This link between ATP utilization and directional movement supports the assertion that NPH-II is a processive molecular motor.

The results suggest the following basic mechanism for RNA duplex unwinding by the DExH RNA helicase NPH-II (Fig. 5): unwinding is preceded by a slow initiation step (2.3 ± 0.2 in Co²⁺, 3.5 ± 0.3 in Mg²⁺). Subsequent unwinding occurs in distinct consecutive steps of roughly one-half helix turn in a defined 3'-to-5' direction with respect to the single-strand 3' overhang. The helicase translocates rapidly along the RNA substrate, unwinding RNA with a rate that is dependent on the identity of the metal ion cofactor (13.4 ± 0.8 min⁻¹ per step for Co²⁺; ≥ 350 min⁻¹ per step for Mg²⁺). During each step, a fraction of protein, predominantly in non-ATP-bound states, dissociates from the RNA. The dependence of reaction on ATP hydrolysis and the direct connection between ATP binding and translocation indicate that NPH-II is a processive, directional motor for unwinding RNA. This activity by DExH/D proteins might ensure processive, regulated rearrangement of structured RNA in macromolecular assemblies during processes such as RNA splicing, export or translation initiation. A rate-limiting step before translocation would provide a straightforward way to control unwinding by blocking and deblocking an early initiation event. Alternatively, factors that control the local concentration of ATP or accessibility of the ATP binding site on the protein could regulate processivity, enabling the helicase to dissociate from RNA at appropriate times. □

Methods

Materials

NPH-II was expressed and purified as described¹⁵. RNAs were prepared by *in vitro* transcription of PCR-generated T7 transcription templates¹⁶ or by chemical synthesis^{17,18}. The PCR-amplified DNA templates represent a segment of the ampicillin-resistance gene of the pBS(+/-) phagemid (Stratagene). To generate perfect blunt-end duplexes, we trimmed the RNA pieces by site-directed processing using engineered DNazymes^{12,19} with 12 nucleotides on each of the binding arms. Bottom-strand RNAs were joined from two separately synthesized oligonucleotides by template-directed ligation using T4 DNA ligase²⁰. RNA duplexes were formed by combining the bottom-strand RNA with a five-fold

bound states is fast compared with translocation with the unwinding rate constant of $k_{U(S)}$ ^{5,7}. k_F is the rate constant for enzyme dissociation from substrate in the ATP-free state, dissociation in the ATP-bound state is not significant (Fig. 4a). k_i is the rate constant for the unwinding initiation step. Reactions were conducted in the presence of trap-RNA (Fig. 1), thus ES dissociation is irreversible.

molar excess of $\gamma^{32}\text{P}$ -labelled top strand in 10 mM MOPS, pH 6.5, 1 mM EDTA. The solution was heated to 95 °C for 2 min and cooled to room temperature over 90 min. Duplexes were separated from single-strand RNA by native PAGE, visualized by radiolytic scanning (Packard Instant Imager) and excised from the gel.

Unwinding reactions

Unwinding reactions were performed at room temperature in 30 μl of 40 mM Tris-HCl buffer, pH 8.0, and 4 mM MgCl_2 or CoCl_2 . The reaction also contained 25 mM NaCl, which was introduced with the protein storage buffer. In a typical reaction, 1–2 nM RNA substrate was incubated with 10–15 nM NPH-II in reaction buffer without ATP at room temperature for 5–7 min. Longer pre-incubation time did not change the reaction kinetics. Saturation of substrate with protein before the reaction was verified by gel-shift analysis¹⁵. The unwinding reaction was initiated by adding the ATP to a final concentration of 3.5 mM, unless otherwise stated. Aliquots at the respective time points were quenched with two volumes of stop buffer (25 mM EDTA, 0.4% SDS, 0.05% BPB, 0.05% XCB, 10% glycerol) containing 200 nM of unlabelled top-strand RNA to prevent re-annealing of unwound duplexes during electrophoresis. Reannealing of unwound duplexes during the reaction did not affect unwinding kinetics under any reaction conditions, as verified by independent measurements of association rate constants of the substrate strands. Trap RNA (200 nM final concentration, added with the ATP) was a single-strand of 39 nucleotides of unrelated sequence. Higher concentrations of trap RNA did not change observed reaction amplitudes.

Unwinding reactions were also monitored in real time using fluorescence energy transfer experiments on substrates labelled at the 3' and 5' ends of the duplex terminus farthest from the single-stranded 3' overhang. These measurements were superimposable with results from gel-shift measurements and confirmed the first-order nature of the time courses as well as the rate constants.

Kinetic analysis

Data were fit numerically using the FITSIM software package²¹. Time courses were corrected for the reaction amplitudes, and initial kinetic parameters were determined using the KINSIM software²². Unwinding step size and rate constants were determined from time courses in Co^{2+} using a kinetic model in which the unwinding reaction consists of N consecutive first-order reactions, where N is the number of kinetic steps. Rate constants (except for the first step, see below) were linked, that is, forced to yield the same value in the fitting. The linked rate constants and the first rate constant were allowed to float simultaneously. On the basis of the observation that an initial step is slower than subsequent steps in Mg^{2+} (see Fig. 3b and text), the first kinetic step was defined as having a different rate constant (k_1) than subsequent kinetic steps (k_U). Using this model, increasing numbers of steps (N) were examined iteratively for each duplex until k_1 and k_U approached constant values for the different duplexes. Convergence occurred when $k_1 = 2.3 \pm 0.2 \text{ min}^{-1}$ and $k_U = 13.4 \pm 0.8 \text{ min}^{-1}$.

Note that k_U is an apparent rate constant that depends on the ATP concentration according to $k_U = k_{U[S]}[\text{ATP}/(\text{ATP} + K'_d)]$, where $k_{U[S]}$ is the respective rate constant at ATP saturation. The step size (m) was calculated according to $m = L/(N - y)$, where y is the number of kinetic steps that do not contribute to the actual strand displacement and L is the duplex length. A step size of $m = 6$ and a value of $y = 1$ were consistently calculated for all duplexes including multi-piece substrates.

The lower limit for the translocation rate constant in Mg^{2+} was estimated by considering that a lag phase was not observed with the longest (83 bp) duplex (Fig. 1b). Assuming an unwinding step size identical to that in Co^{2+} (6 bp), and that k_U must be equal or greater than a value that does not cause an apparent lag phase in 14 unwinding steps, the translocation rate constant was estimated as $k_U > \sim 350 \text{ min}^{-1}$.

Equation (1) was derived from the expression describing the dependence of reaction amplitude (A) on the processivity (P)¹³:

$$A = (1 - x) P^{1/m} \quad (2)$$

where x is the fraction of protein that dissociates from the substrate before unwinding occurs. P is then substituted with a term that relates processivity and ATP concentration under the conditions specified by the scheme in Fig. 5.

$$P = \frac{k_{U[S]} \left(\frac{[\text{ATP}]}{K'_d + [\text{ATP}]} \right)}{k_{U[S]} \left(\frac{[\text{ATP}]}{K'_d + [\text{ATP}]} \right) + k_f \left(1 - \frac{[\text{ATP}]}{K'_d + [\text{ATP}]} \right)} \quad (3)$$

where K'_d is the apparent dissociation constant of ATP from the enzyme substrate complex.

Received 29 July; accepted 28 October 1999.

- de la Cruz, J., Kressler, D. & Linder, P. Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. *Trends Biochem. Sci.* **24**, 192–198 (1999).
- Kadare, G. & Haenni, A. L. Virus encoded RNA helicases. *J. Virol.* **71**, 2583–2590 (1997).
- Wagner, J. D. O., Jankowsky, E., Company, M., Pyle, A. M. & Abelson, J. N. The DEAH-box protein Prp22 is an ATPase that mediates ATP dependent mRNA release from the spliceosome and unwinds RNA duplexes. *EMBO J.* **17**, 2926–2937 (1998).
- Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors clocks springs and things. *Cell* **90**, 1041–1050 (1998).
- Shuman, S. Vaccinia Virus RNA helicase: an essential enzyme related to the DE-H family of RNA dependent NTPases. *Proc. Natl Acad. Sci. USA* **89**, 10935–10939 (1992).

- Shuman, S. Vaccinia Virus RNA helicase. *J. Biol. Chem.* **268**, 11798–11802 (1993).
- Gross, C. H. & Shuman, S. Vaccinia virus RNA helicase: nucleic acid specificity in duplex unwinding. *J. Virol.* **70**, 2615–2619 (1996).
- Ali, J. A. & Lohman, T. M. Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. *Science* **275**, 377–380 (1997).
- Shimamoto, N. One-dimensional diffusion of proteins along DNA. *J. Biol. Chem.* **274**, 15239–15296 (1999).
- Kellemen, B. R. & Raines, R. T. Extending the limits to enzymatic catalysis: diffusion of ribonuclease A in one dimension. *Biochemistry* **27**, 5302–5307 (1999).
- Jeltsch, A. & Pingoud, A. Kinetic characterization of linear diffusion of the restriction endonuclease EcoRV on DNA. *Biochemistry* **24**, 2160–2169 (1998).
- Santoro, S. W. & Joyce, G. F. A general purpose RNA-cleaving DNA enzyme. *Proc. Natl Acad. Sci. USA* **94**, 4262–4266 (1997).
- Lohman, T. M. & Bjoernson, K. P. Mechanisms of helicase catalyzed DNA unwinding. *Annu. Rev. Biochem.* **65**, 169–214 (1996).
- Schnapp, B. Molecular Motors. Two heads are better than one. *Nature* **373**, 655–666 (1995).
- Gross, C. H. & Shuman, S. Mutational analysis of vaccinia virus nucleoside triphosphate hydrolyase II, a DExH box RNA helicase. *J. Virol.* **69**, 4727–4736 (1995).
- Jankowsky, E. & Schwenzer, B. Efficient improvement of hammerhead ribozyme mediated cleavage of long substrates by oligonucleotide facilitators. *Biochemistry* **37**, 15313–15321 (1996).
- Xiang, Q., Qin, P., Michels, W. J., Freeland, K. & Pyle, A. M. Sequence specificity of a group II intron ribozyme: multiple mechanisms for promoting unusually high discrimination against mismatched targets. *Biochemistry* **37**, 3839–3849 (1998).
- Wincott, F. et al. Synthesis, deprotection, analysis and purification of RNA and ribozymes. *Nucleic Acids Res.* **23**, 2677–2684 (1995).
- Pyle, A. M., Chu, V. T., Jankowsky, E. & Boudvillain, M. Using DNazymes to cut, process and map RNA molecules for structural studies or modification. *Methods Enzymol.* **317**, 140–146 (2000).
- Moore, M. J. & Sharp, P. A. Site specific modification of pre-messenger RNA—the 2'-hydroxyl groups at the splice sites. *Science* **256**, 992–997 (1992).
- Zimmerle, C. & Frieden, C. Analysis of progress curves by simulations generated by numerical integration. *Biochem. J.* **258**, 381–387 (1989).
- Barshop, B. A. & Frieden, C. Analysis of numerical methods for computer simulation of kinetic processes: development of KINSIM—a flexible portable system. *Anal. Biochem.* **130**, 134–145 (1983).

Acknowledgements

We thank the members of the Pyle laboratory, particularly M. Boudvillain, J. B. Green and P. Z. Qin for many insightful discussions. E.J. was supported by a Curt-Engelhorn postdoctoral fellowship from the German Center for Cancer Research. A.M.P. is an assistant investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to A.M.P.

DNA-bound structures and mutants reveal abasic DNA binding by APE1 DNA repair and coordination

Clifford D. Mol*, Tadahide Izumi†, Sankar Mitra† & John A. Tainer*

* Skaggs Institute for Chemical Biology, and the Department of Molecular Biology, MB-4, The Scripps Research Institute, 10550 North Torrey Pines Rd., La Jolla, California 92037-1027, USA

† Sealy Center for Molecular Science, Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas 77555-1079, USA

Non-coding apurinic/apyrimidinic (AP) sites in DNA are continually created in cells both spontaneously and by damage-specific DNA glycosylases¹. The biologically critical human base excision repair enzyme APE1 cleaves the DNA sugar-phosphate backbone at a position 5' of AP sites to prime DNA repair synthesis^{2–4}. Here we report three co-crystal structures of human APE1 bound to abasic DNA which show that APE1 uses a rigid, pre-formed, positively charged surface to kink the DNA helix and engulf the AP-DNA strand. APE1 inserts loops into both the DNA major and minor grooves and binds a flipped-out AP site in a pocket that excludes DNA bases and racemized β -anomer AP sites. Both the APE1 active-site geometry and a complex with cleaved AP-DNA and Mn^{2+} support a testable structure-based catalytic mechanism. Alanine substitutions of the residues that penetrate the DNA helix unexpectedly show that human APE1 is

structurally optimized to retain the cleaved DNA product. These structural and mutational results show how APE1 probably displaces bound glycosylases and retains the nicked DNA product, suggesting that APE1 acts *in vivo* to coordinate the orderly transfer of unstable DNA damage intermediates between the excision and synthesis steps of DNA repair.

Despite the essential and central role of APE1 in human base excision repair (BER)^{2,5}, there have been no clear hypotheses concerning how APE1 coordinates human DNA repair between the first damage-specific stage carried out by distinct DNA glycosylases, and the subsequent damage-general stage of DNA repair synthesis. To resolve the molecular basis for substrate recognition, catalysis and progression of BER steps coordinated by APE1, we determined two co-crystal structures of human APE1 bound to 11-base-pair (bp) and 15-bp DNA substrates with synthetic abasic sites, at 2.95 Å and 2.65 Å resolution, respectively. We also determined the 3.0 Å resolution ternary product complex of APE1, nicked abasic DNA and a divalent metal ion (Fig. 1a, b; see Methods).

APE1 electrostatically orients a rigid, pre-formed DNA-binding face and penetrates the DNA helix from both the major and minor grooves, stabilizing an extrahelical conformation for target abasic

nucleotides and excluding normal DNA nucleotide and racemized abasic site binding (Fig. 2a, b, and see animation in Supplementary information). Although nucleotide flipping to expose damaged bases is characteristic of DNA glycosylases, it is not obviously required by APE1 and other DNA-backbone-cleaving endonucleases, and is not seen for the very short patch DNA repair endonuclease (VSR)⁶. However, abasic nucleotide flipping also occurs in a structurally unrelated AP endonuclease, endonuclease IV (ref. 7), resembling that seen in these APE1–DNA complexes, and suggesting that nucleotide flipping may be a property of AP endonucleases.

Unlike the open-to-closed conformational transition that occurs in DNA glycosylases and polymerases, the APE1 structure is pre-formed for abasic DNA recognition. The r.m.s. deviation is only ~0.7 Å between Cα atoms of the free APE1 enzyme⁸ and those of DNA-bound structures. The overall DNA-binding position resembles the model suggested by the unbound enzyme⁸; however, the APE1-bound DNA is severely distorted, with the DNA bent ~35° and the helical axis kinked or displaced ~5 Å, where the two inserted loops cap the extrahelical AP site (Fig. 2a, b). In agreement with biochemical results⁹, the enzyme–DNA interface encompasses both DNA strands but is dominated by contacts with the AP–DNA strand.

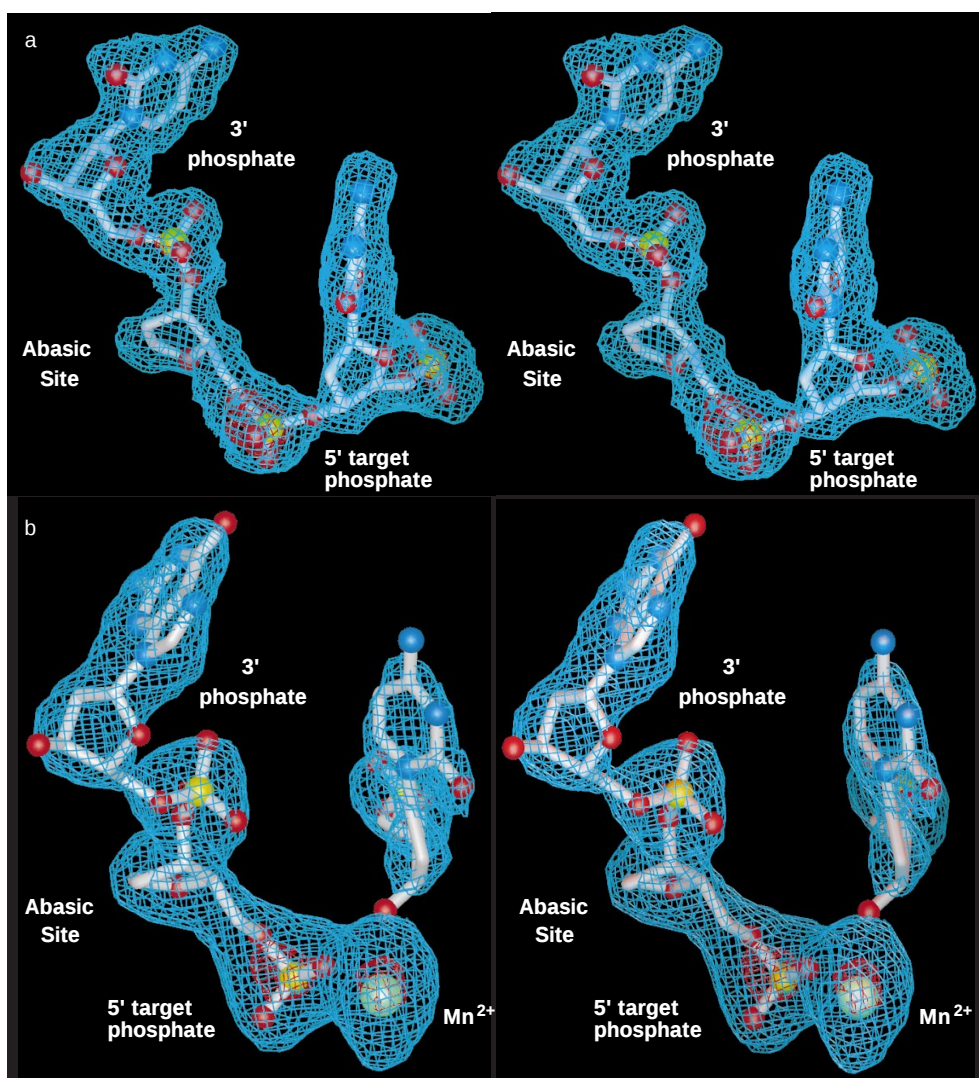


Figure 1 Stereo views of simulated-annealed omit electron density for APE1–DNA complexes showing the flipped-out abasic site and phosphodiester bond cleavage. Electron-density maps were calculated excluding the abasic and flanking nucleotides, and the divalent metal ion in **b**. APE1-bound DNA is shown with phosphorus (yellow), oxygen (red), nitrogen (blue) and carbon polytubes (white) highlighted. **a**, The APE1 active site

from the 2.65 Å resolution APE1–DNA complex. Omit electron-density contours are shown at 3σ (blue) and 9σ (red). **b**, The APE1 active site from the ternary complex with bound Mn²⁺ (green sphere) and nicked DNA. Omit electron-density contours are shown at 4σ (blue) and 10σ (red).

The positively charged, grooved APE1 DNA-binding face (Fig. 2b) buries nearly $1,600\text{\AA}^2$ of DNA solvent-accessible surface area, with $\sim 1,000\text{\AA}^2$ contributed by the AP-DNA strand, and $\sim 600\text{\AA}^2$ by the opposing strand. The interface is centred on the flipped-out abasic nucleotide and its flanking phosphates, which account for $\sim 500\text{\AA}^2$ of the total buried DNA surface area.

APE1 stabilization of the kinked abasic DNA is mediated by residues emanating from four loops and from one α -helix (Fig. 2a). As AP sites fit within a B-DNA helix in solution¹⁰, APE1 DNA contacts facilitate and stabilize the extrahelical AP site. 5' of the AP site, three APE1 helical residues (Arg 73, Ala 74 and Lys 78) contact three consecutive DNA phosphates of the opposing strand, and Tyr 128 and Gly 127 span and widen the minor groove $\sim 2\text{\AA}$. These 5' contacts may anchor the DNA for the extreme kinking caused by three APE1 loops 3' of the extrahelical abasic site (Fig. 2a). Immediately 3' of the AP site, APE1 binds two phosphates and braces the AP-DNA backbone for double-loop insertion. Across

from the flipped-out AP site, Met 270 inserts through the minor groove to pack against the orphan base partner of the abasic site and to occupy the space where it would be in unknicked B-DNA. The APE1-bound DNA is so severely distorted, however, that the orphan base still stacks with the 5' base (Fig. 2a). Above the abasic site, Arg 177 inserts through the DNA major groove and delivers a hydrogen bond to the AP site 3' phosphate (Fig. 3a). This DNA major groove interaction is unusual for a DNA repair enzyme, and, as the sequence and conformation of the Arg 177 loop is unique to APE1, it probably reflects specific APE1 functions. The double-loop penetration of the DNA helix, from the insertion of Arg 177 in the major groove and Met 270 in the minor groove, caps the DNA-bound enzyme active site, stabilizes the extrahelical AP site conformation and effectively locks APE1 onto the AP-DNA.

Our APE1 substrate and product complexes suggest a reaction mechanism distinct from previous predictions^{8,11,12}, in which the target phosphate is polarized about the scissile bond and attacked by

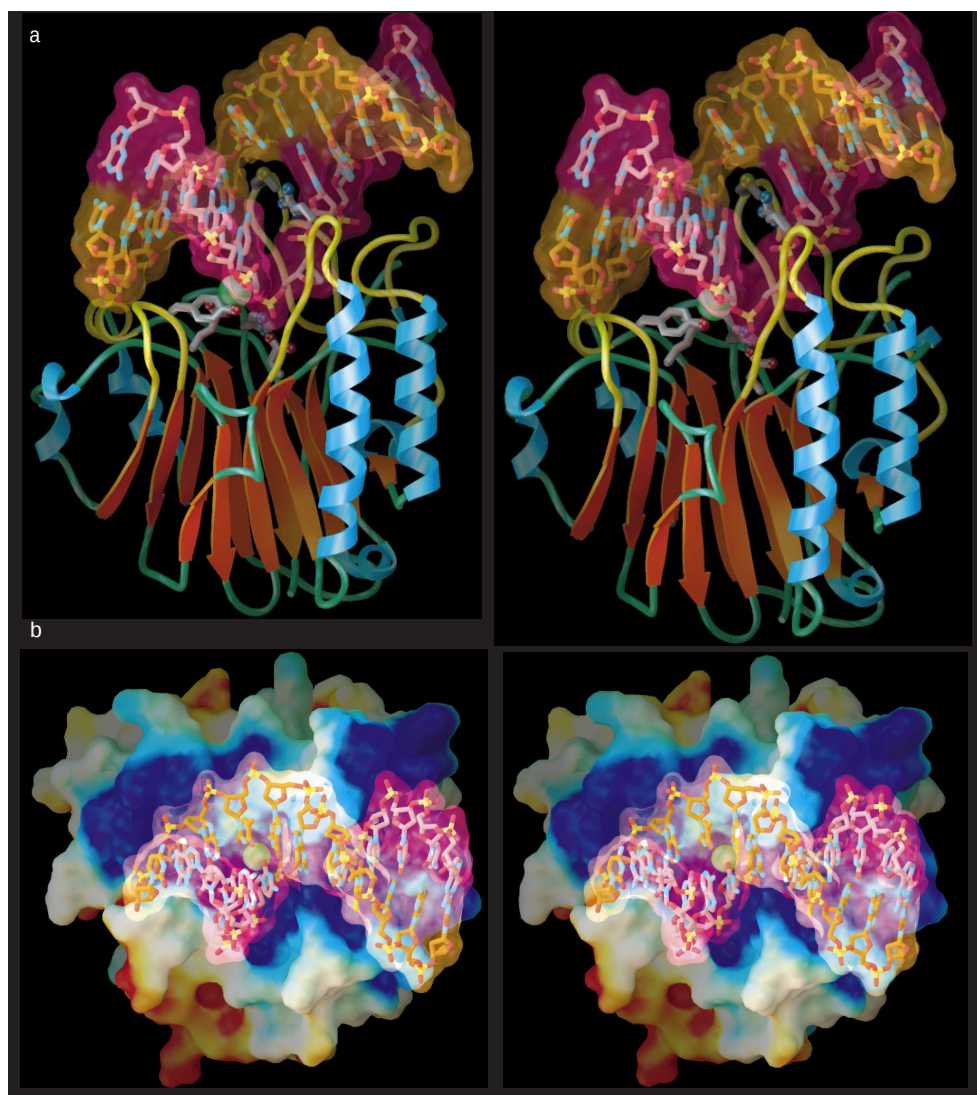


Figure 2 Stereo views of the APE1 complex with AP-DNA showing double-loop penetration of the DNA helix, plus charge and surface complementarity. **a**, DNA (top) binding to APE1 (bottom) from the co-crystal structure of APE1 bound to the 11-bp abasic DNA. The complex is viewed from the side that is roughly perpendicular to the kinked DNA helix axis oriented with the AP strand running 5' (left) to 3' (right), and shows the two central, six-stranded APE1 β -sheets (orange arrows), α - and 3_{10} -helices (blue coils), and coils (green tubes). Five APE1 DNA-binding regions (yellow; residues 73–79, 126–129, 174–178, 221–228 and 268–272) contact both the AP-DNA strand (pink carbon polytubes and transparent surface) and the opposing strand (orange polytubes and

surface). The position of the divalent metal ion (green sphere) which is 5' of the AP site as derived from the ternary product complex, and the side chains for key residues are shown. **b**, DNA binding to APE1 viewed looking down on the enzyme–DNA interface, rotated $\sim 90^\circ$ from the view in **a**. Arg 177 penetrates the DNA major groove (bottom) and Met 270 inserts through the DNA minor groove (top), capping the DNA-bound APE1 active site. The APE1-bound DNA, which is bent $\sim 35^\circ$ with the helix axis kinked $\sim 5\text{\AA}$, complements the positively charged DNA-binding surface of APE1, coloured red (-1.5 kT e^{-1}) to blue ($+1.5\text{ kT e}^{-1}$) according to its electrostatic potential (calculated including the Mn^{2+} ion), which engulfs the AP-DNA strand (pink) around the flipped-out AP site.

an aspartate-activated hydroxyl nucleophile, with the metal ion stabilizing the transition state and the leaving group of the phosphodiester cleavage reaction. On the 3' side of the APE1–DNA complex, Trp 280 contacts a DNA phosphate, and a chain of three asparagine residues, Asn 222, Asn 226 and Asn 229, lines the back of the active site and contacts two 3' AP-DNA phosphates. On the 5' side of the active site, the side chains of Asn 174 and Asn 212 form hydrogen bonds with O5' and O2P of the target phosphate, and the metal ion, bound by Glu 96, ligates the O3' atom (Fig. 3a). In our structure-based reaction scheme, which agrees with known mutagenesis results^{12,13}, the conserved Asp 283/His 309 pair makes a direct hydrogen bond from His 309 Nε2 to the remaining O1P atom (Fig. 3b, and animation in Supplementary information). These APE1–phosphate contacts orient and polarize the scissile P–O3' bond with the Asp 210 side chain favourably aligned to activate an attacking nucleophile. Positioned by hydrogen bonds with the backbone amide of Asn 212 and Asn 68 Nδ2, the pK_a of buried Asp 210 is probably elevated, and along with the Asp 283/His 309 phosphate interaction, probably results in the observed maximal APE1 activity at pH 7.5, with a pH 6.7–9 activity profile¹⁴. Protonation of Asp 210 can generate the proposed attacking hydroxyl nucleophile (Fig. 3b). The divalent metal ion may facilitate the O3' leaving group, either by direct ligation or through a water in the first hydration shell of a Mg²⁺ ion. Specific binding of APE1 to extrahelical AP sites derives from a hydrophobic pocket

bordered by Phe 266, Trp 280 and Leu 282, which pack with the hydrophobic side of the abasic deoxyribose (Fig. 3a). This tight packing excludes the binding of normal DNA nucleotides, and is specific for the α-anomer of natural AP sites generated as the direct product of DNA glycosylases and not the β-anomer obtained by racemization of AP sites in solution. Thus, this APE1 specificity indicates that *in vivo* the AP site may be passed directly to APE1 from the damage-specific DNA glycosylases, thereby avoiding the release of repair intermediates more toxic than the original damage.

To investigate the role of double-loop penetration in AP-site detection and APE1 function, we determined the effects on enzyme activity of site-directed alanine substitutions for Met 270, Met 271 and Arg 177. Unexpectedly, none of these residues is required to promote abasic nucleotide flipping (Fig. 4), indicating that APE1 may be a structure-specific nuclease that detects and productively binds DNA that can adopt the kinked conformation and can present a flipped-out AP site to the APE1 recognition pocket. Both the Met270Ala and Met271Ala mutant enzymes have essentially wild-type levels of activity, whereas the specific activity is, surprisingly, increased for the Arg177Ala mutant enzyme (Fig. 4). Under the appropriate assay conditions of excess substrate, these data support the hypothesis that the Arg 177 intercalation and phosphate interaction slow APE1 dissociation from the cleaved product¹⁵. Moreover, kinetic analysis of the Arg177Ala mutant confirms that it is

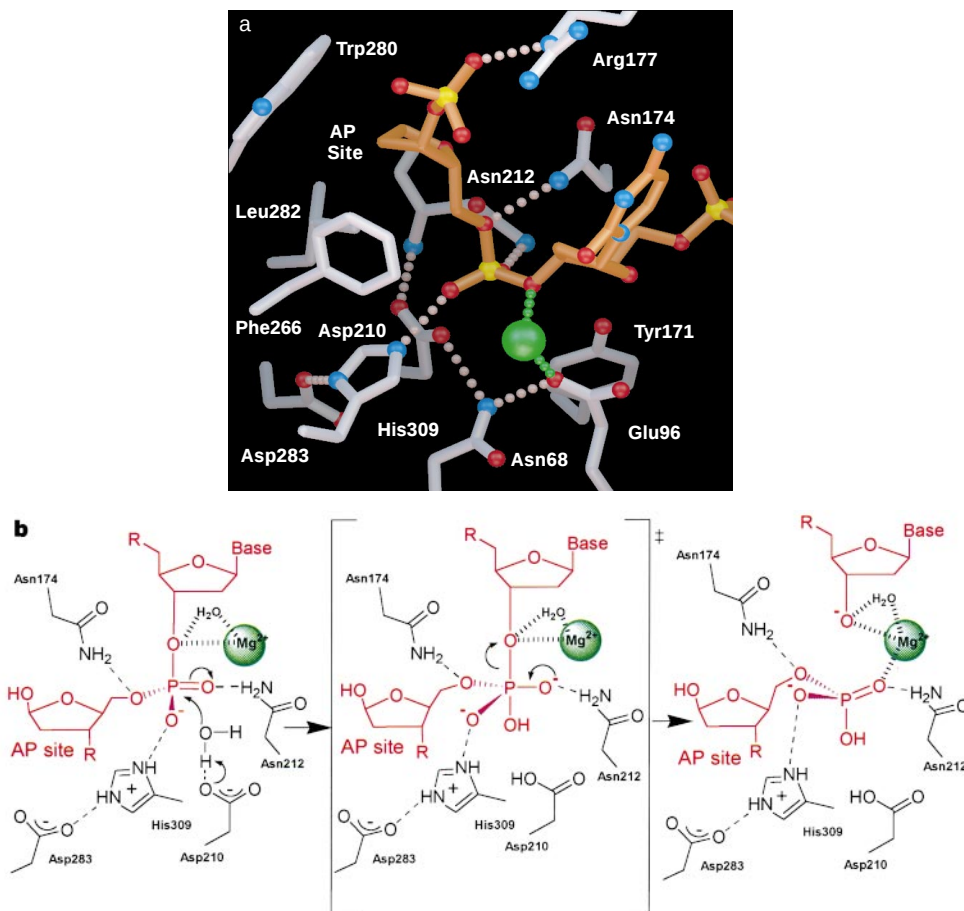


Figure 3 APE1 interactions with the flipped-out AP site provide damage specificity and suggest a specific reaction mechanism for phosphodiester bond cleavage. **a**, APE1 active-site interactions with the flipped-out abasic deoxyribose and target 5' phosphate. The hydrophobic face of the extrahelical AP site packs within a complementary APE1 pocket formed by the side chains of Phe 166, Trp 280 and Leu 282. The Arg 177 side chain inserts through the kinked DNA major groove to form a hydrogen bond to the non-target AP site 3' phosphate. The target AP site 5' phosphate is oriented by hydrogen bonds with Asn 174, Asn 212 and His 309. Asp 210 is oriented by hydrogen bonds from the Asn 212 backbone amide and Asn 68 Nδ2. The metal ion position (derived from the

ternary product complex) stabilizes the transition state and facilitates the O3' leaving group, perhaps through a first hydration shell water molecule of a Mg²⁺ ion. **b**, Structure-based reaction mechanism for phosphodiester bond cleavage. Substrate AP-DNA is oriented by the bound divalent metal ion and APE1 active-site residues for attack of an hydroxyl nucleophile activated by buried Asp 210 (left panel). Collapse of the pentacoordinate transition state (middle panel) leads to cleavage of the scissile P–O3' bond, with the transition state and O3' leaving group stabilized by the metal ion, leading to bond cleavage and inversion of the phosphate configuration (right panel).

over 25% more efficient than wild-type APE1, with an increased Michaelis constant (K_m) offset by a larger increase in the rate constant (k_{cat}) (Fig. 4). As evolutionary selection occurs at the level of biological function rather than single enzyme rates, this apparent dichotomy supports a biological role for tight product binding by APE1. Thus, the biological need to coordinate APE1 with downstream DNA BER enzymes^{16,17} has evidently guided the evolution of the APE1 structure, rather than maximizing catalytic efficiency and allowing rapid release of toxic, incised DNA products.

Together these APE1–DNA complex structures and mutagenesis results suggest a coherent and testable model for the orderly transfer of DNA damage in BER. First, we propose that the extensive double-strand DNA (dsDNA)-binding interaction surface of APE1 and pronounced kinking of the DNA helix as compared with DNA glycosylases allow APE1 to stage human DNA repair by displacing bound glycosylases from AP sites, consistent with the observed APE1 enhancements of human uracil- and thymine-DNA glycosylase activities^{18,19}. Second, by being optimized to retain its nicked DNA product and to also interact with the repair polymerase Pol β in the presence of DNA, the kinked APE1–DNA complex may recruit Pol β to sites of DNA damage. The high relative amounts of APE1 per cell (350,000 to 7,000,000)²⁰ compared with Pol β (~50,000 molecules)²¹ suggest that functionally relevant interactions with Pol β may be mediated by recognition of the distorted APE1–DNA complex. Through the known interactions^{22,23} of Pol β , XRCC1 and DNA ligase III, the entire DNA repair synthesis machinery may then be assembled. Most DNA BER synthesis replaces only the single damaged nucleotide, and the religated and repaired DNA should disfavor the ~90° bend for Pol β binding, and promote enzyme dissociation after repair synthesis. Thus, these structural and mutational analyses of human APE1 reveal adaptations that are counterintuitive in terms of single enzyme catalysis and provide insights into the evolutionary optimization of APE1 biological function for maintaining DNA integrity through the efficient transfer of DNA damage in BER. □

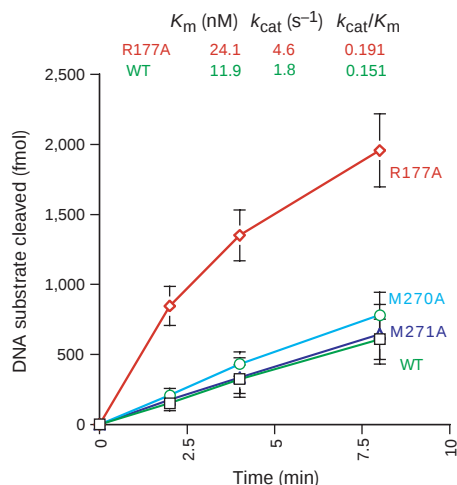


Figure 4 Activities and kinetics of wild-type and mutant human APE1 enzymes. AP endonuclease activities of wild-type APE1 (green), and Met 270 (cyan), Met 271 (blue) and Arg 177 (red) site-directed alanine mutant enzymes. Enzyme activity was assayed using a 43-bp oligonucleotide (160 nM) containing a single AP site analogue⁴ with a [³²P]ATP 5' end label and 100 pg of purified protein. Wild-type APE1 and mutant enzymes were incubated at 37° with ~160 nM substrate in a volume of 20 μ l containing 60 mM Tris pH 8.0, 1 mM MgCl₂, 0.2 mM EDTA, 1 mM DTT and 100 μ g ml⁻¹ bovine serum albumin. The reaction was stopped at the indicated time points, separated on 20% polyacrylamide gels containing 7 M urea, and quantified by phosphorimage analyses. Results are the average of three independent experiments with ranges as indicated by the error bars. Enzyme kinetic analyses of Arg177Ala were performed as described for wild-type APE1 (ref. 13).

Methods

Biochemistry and crystallization

Expression and purification of full-length APE1 and of a fully active, N-terminal truncated enzyme containing residues 40–318 (NΔ40APE1) were done as described¹³, with expression of the NΔ40 APE1 gene in *E. coli* TZ1 (xth nfo DE3) to create an N-terminal 6xHis tag. The enzyme was treated with thrombin (1:1,000 for 6 h at 37°C) to remove the 6xHis tag; and purified recombinant human NΔ40APE1 enzyme was concentrated to 20 mg ml⁻¹, and mixed with an equal volume of ~2 mM 11-bp dsDNA (5'-GCTACGATCG-3') containing a synthetic tetrahydrofuran (F) AP site analogue (Midland Certified Reagent Co.) This AP site analogue is more stable than, and binds to and is cleaved by APE1 at rates comparable to, a deoxyribose with a C1' hydroxyl group^{3,12,24,25}.

Co-crystals were grown by mixing equal volumes of enzyme–DNA solution with reservoir solutions of 20% polyethylene glycol monomethylether (MPEG) 5000 (Sigma), 50 mM 2-(N-morpholino)ethanesulphonic acid buffer pH 6.5 and 200 mM Li₂SO₄. Crystals grew to full size (0.5 × 0.2 × 0.06 mm) in ~2 weeks, belong to the body-centred, tetragonal space group *I*₄, *a*=*b*=123.4, *c*=107.1 Å, and contain two APE1–DNA complexes in the asymmetric unit. X-ray diffraction data were collected from single crystals flash-cooled at -170° at beamline 9-1 of the Stanford Synchrotron Radiation Laboratory (SSRL) on a 34.5 cm MARresearch imaging plate and processed with DENZO and SCALEPACK²⁶ to yield 42,400 observations of 15,572 unique reflections to 2.95 Å resolution (93% complete) with an average *I*/ σ (*I*) of 12.7 and an overall *R*_{sym} between symmetry related reflections of 0.050 (94% complete, *I*/ σ (*I*) = 2.2, *R*_{sym} = 0.302; for the 2.95 to 3.06 Å resolution shell). The APE1–DNA–Mn²⁺ complex crystals were obtained by co-crystallization with 25 mM MnCl₂ and the same 11-bp dsDNA, and belong to space group *P*₂₁2₁2₁, *a*=90.1, *b*=98.3, *c*=101.1 Å, with 41,650 observations of 16,996 unique reflections to 3.0 Å (91% complete), an overall *I*/ σ (*I*) of 8.6, and *R*_{sym} = 0.086 (92% complete, *I*/ σ (*I*) = 2.3, *R*_{sym} = 0.424; for the 3.0–3.11 Å resolution shell). Crystals of APE1 bound to a 15-bp dsDNA oligonucleotide (5'-GCGTCCFCGACGACG-3') are grown using full-length, wild-type APE1 from well solutions containing 100 mM Cacodylate (pH 6.5), 200 mM Li₂SO₄ and 27% MPEG 2000 (Sigma). These crystals also have two APE1–DNA complexes in the asymmetric unit, and belong to space group *P*₂₁, *a*=71.2, *b*=72.4, *c*=93.7 Å, β =94.3, with 61,212 observations of 24,575 unique reflections to 2.65 Å (93% complete), *I*/ σ (*I*) = 7.8, and *R*_{sym} = 0.087 (92% complete, *I*/ σ (*I*) = 2.2, *R*_{sym} = 0.367; for the 2.65–2.75 Å resolution shell).

Structure solution and refinement

The 2.95 Å resolution APE1–DNA structure was solved by molecular replacement with AMoRe²⁷ using free APE1 (ref. 8) as a search model and all data from 4.0–15.0 Å resolution. The correct solutions gave a correlation coefficient of 0.36 and an *R* value of 0.47, with the corresponding values for the next highest peak 0.25 and 0.51. This model was refined with X-PLOR 3.8 (ref. 28) and electron density for the DNA was clearly visible in σ A-weighted²⁹ 2F_o–F_c and simulated-annealed omit electron-density maps. Cycles of model building and stereochemically restrained simulated annealing and positional refinement allowed all the DNA to be placed, with manual inspection and rebuilding of the structure performed with XFIT³⁰. The APE1–DNA complex structure contains 4,364 non-hydrogen protein atoms and 874 DNA atoms and has been refined to an *R* value of 0.192 (*R*_{free} = 0.308) for 12,453 reflections from 20.0 to 2.95 Å resolution with *F* > 3.0 σ F (74% complete), with r.m.s. deviations from ideality of 0.010 Å for bond lengths, and 1.48° for bond angles. This structure was used to solve the ternary product complex with AMoRe (correlation coefficient = 0.60, *R* = 0.37), which was refined as with the APE1–DNA complex and data from 20.0 to 3.0 Å resolution (13,766 reflections, 75% complete), to an *R* = 0.181 (*R*_{free} = 0.274), and r.m.s. deviations of 0.006 Å for bond lengths and 1.48° for bond angles. The 2.65 Å resolution APE1–DNA complex structure was also solved with AMoRe and refined with X-PLOR 3.8 using data from 20.0 to 2.65 Å with *F* > 2.0 σ F (25,211 reflections, 90% complete) to yield an *R* value of 0.203 (*R*_{free} = 0.291) with 0.007 Å deviation in bond lengths and 1.40° for bond angles. The ~40 N-terminal residues and the last base pair are disordered in this structure, and all three structures contain a *cis* peptide bond between Val 247 and Pro 248.

Received 23 September; accepted 15 November 1999.

- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Wilson, D. M. & Thompson, L. H. Life without DNA repair. *Proc. Natl Acad. Sci. USA* **94**, 12754–12757 (1997).
- Wilson, D. M., Takeshita, M., Grollman, A. P. & Demple, B. Incision activity of human apurinic endonuclease (Ape) at abasic site analogs in DNA. *J. Biol. Chem.* **270**, 16002–16007 (1995).
- Izumi, T. & Mitra, S. Deletion analysis of human AP-endonuclease: minimum sequence required for the endonuclease activity. *Carcinogenesis* **19**, 525–527 (1998).
- Lindahl, T. & Wood, R. D. Quality control in DNA repair. *Science* **286**, 1897–1905 (1999).
- Tsutakawa, S. E., Jingami, H. & Morikawa, K. Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell* **99**, 615–623 (1999).
- Hosfield, D. J., Guan, Y., Haas, B. J., Cunningham, R. P. & Tainer, J. A. Structure of the DNA repair enzyme endonuclease IV and its DNA complex: double-nucleotide flipping at abasic sites and three-metal-ion catalysis. *Cell* **98**, 397–408 (1999).
- Gorman, M. A. et al. The crystal structure of the human DNA repair endonuclease HAP1 suggests the recognition of extra-helical deoxyribose at DNA abasic sites. *EMBO J.* **16**, 6548–6558 (1997).
- Wilson, D. M., Takeshita, M. & Demple, B. Abasic site binding by the human apurinic endonuclease, Ape, and determination of the DNA contact sites. *Nucleic Acids Res.* **25**, 933–939 (1997).
- Withika, J. M., Wilde, J. A. & Bolton, P. H. Characterization of conformational features of DNA heteroduplexes containing aldehydic abasic sites. *Biochemistry* **30**, 9931–9940 (1991).
- Mol, C. D., Kuo, C.-F., Thayer, M. M., Cunningham, R. P. & Tainer, J. A. Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* **374**, 381–386 (1995).
- Erzberger, J. P. & Wilson, D. M. The role of Mg²⁺ and specific amino acid residues in the catalytic

- reaction of the major human abasic endonuclease: new insights from EDTA-resistant incision of acyclic abasic site analogs and site-directed mutagenesis. *J. Mol. Biol.* **290**, 447–457 (1999).
13. Izumi, T. *et al.* Intragenic suppression of an active site mutation in the human apurinic/aprimidinic endonuclease. *J. Mol. Biol.* **287**, 47–57 (1999).
 14. Kane, C. M. & Linn, S. Purification and characterization of an apurinic/aprimidinic endonuclease from HeLa cells. *J. Biol. Chem.* **256**, 3405–3414 (1981).
 15. Masuda, Y., Bennett, R. A. & Dimple, B. Dynamics of the interaction of human apurinic endonuclease (Ape1) with its substrate and product. *J. Biol. Chem.* **273**, 30352–30359 (1998).
 16. Bennett, R. A. O., Wilson, D. M., Wong, D. & Dimple, B. Interaction of human apurinic endonuclease and DNA polymerase β in the base excision repair pathway. *Proc. Natl Acad. Sci. USA* **94**, 7166–7169 (1997).
 17. Prasad, R. *et al.* Specific interaction of DNA polymerase β and DNA ligase I in a multiprotein base excision repair complex from bovine testis. *J. Biol. Chem.* **271**, 16000–16007 (1996).
 18. Parikh, S. S. *et al.* Base-excision repair initiation revealed by crystal structures and DNA-binding kinetics of human uracil-DNA glycosylase bound to DNA. *EMBO J.* **17**, 5214–5226 (1998).
 19. Waters, T. R., Gallinari, P., Jiricny, J. & Swann, P. F. Human thymine DNA glycosylase binds to apurinic sites in DNA but is displaced by human apurinic endonuclease 1. *J. Biol. Chem.* **274**, 67–74 (1999).
 20. Chen, D. S., Herman, V. & Dimple, B. Two distinct human DNA diesterases that hydrolyze 3'-blocking deoxyribose fragments from oxidized DNA. *Nucleic Acids Res.* **19**, 5907–5914 (1991).
 21. Horton, J. K., Srivastava, D. K., Zmudzka, B. Z. & Wilson, S. H. Strategic down-regulation of DNA polymerase beta by antisense RNA sensitizes mammalian cells to specific DNA damaging agents. *Nucleic Acids Res.* **23**, 3810–3815 (1995).
 22. Kubota, Y. *et al.* Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase β and the XRCC1 protein. *EMBO J.* **15**, 6662–6670 (1996).
 23. Caldecott, K. W., McKeown, C. K., Tucker, J. D., Ljungquist, S. & Thompson, L. H. An interaction between the mammalian DNA repair protein XRCC1 and DNA ligase III. *Mol. Cell. Biol.* **14**, 68–76 (1994).
 24. Takeshita, M., Chang, C. N., Johnson, E., Will, S. & Grollman, A. P. Oligodeoxynucleotides containing synthetic abasic sites. Model substrates for DNA polymerases and apurinic/aprimidinic endonucleases. *J. Biol. Chem.* **262**, 10171–10179 (1987).
 25. Erzberger, J. P., Barsky, D., Scharer, O. D., Colvin, M. E. & Wilson, D. M. Elements in abasic site recognition by the major human and *Escherichia coli* apurinic/aprimidinic endonucleases. *Nucleic Acids Res.* **26**, 2771–2778 (1998).
 26. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–325 (1997).
 27. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
 28. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallographic R factor refinement by molecular dynamics. *Science* **235**, 458–460 (1987).
 29. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140–149 (1986).
 30. McRee, D. E. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>).

Acknowledgements

We thank S. S. Parikh, C. D. Putnam, D. S. Daniels and D. J. Hosfield for helpful discussions, and the staff and facilities at SSRL. This work was supported by the NIH, by a Laboratory Directed Research and Development grant from the Lawrence Berkeley Laboratory, by a cancer research supplement administered through LBNL from the National Cancer Institute (to J.A.T., S.M. and P. Cooper), the Skaggs Institute for Chemical Biology and a Special Fellowship from the Leukemia Society of America (to C.D.M.).

Correspondence and requests for materials should be addressed to J.A.T. (e-mail: jat@scripps.edu). Coordinates have been deposited in the RCSB Protein Data Bank under accession codes 1DE8, 1DE9 and 1DEW.

Design of single-layer β -sheets without a hydrophobic core

Shohei Koide*, Xiaolin Huang*, Karl Link*, Akiko Koide*, Zimei Bu† & Donald M. Engelman†

* Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, New York 14642, USA

† Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06511, USA

The hydrophobic effect is the main thermodynamic driving force in the folding of water-soluble proteins^{1,2}. Exclusion of nonpolar moieties from aqueous solvent results in the formation of a hydrophobic core in a protein, which has been generally considered essential for specifying and stabilizing the folded structures of proteins^{1–6}. Outer surface protein A (OspA) from *Borrelia*

burgdorferi contains a three-stranded β -sheet segment which connects two globular domains⁷. Although this single-layer β -sheet segment is exposed to solvent on both faces and thus does not contain a hydrophobic core, the segment has a high conformational stability⁸. Here we report the engineering of OspA variants that contain larger single-layer β -sheets (comprising five and seven β -strands) by duplicating a β -hairpin unit within the β -sheet. Nuclear magnetic resonance and small-angle X-ray scattering analyses reveal that these extended single-layer β -sheets are formed as designed, and amide hydrogen-deuterium exchange and chemical denaturation show that they are stable. Thus, interactions within the β -hairpin unit and those between adjacent units, which do not involve the formation of a hydrophobic core, are sufficient to specify and stabilize the single-layer β -sheet structure. Our results provide an expanded view of protein folding, misfolding and design.

The crystal structure of OspA revealed that this protein has an unusual architecture⁷: it is dumbbell-shaped and contains a three-stranded 'single-layer' β -sheet in the centre (Fig. 1). The single-layer β -sheet segment is rich in polar amino acids, and its amino-acid sequence does not follow the canonical, alternating hydrophobic/hydrophilic pattern usually found in amphipathic β -sheets (Fig. 2). Our previous studies showed that the solution structure of OspA is similar to the crystal structure^{9,10}, and that the protein, including the unusual single-layer β -sheet region, is highly stable^{8,11}.

The OspA single-layer β -sheet, although it does not contain a hydrophobic core, buries nonpolar surfaces to a degree comparable to that found in small globular proteins that do contain a hydrophobic core⁸. We attributed this effective burial of nonpolar surfaces to strategically placed amino acids with long side chains in the β -sheet. Despite the lack of a hydrophobic core in this region, the single-layer β -sheet architecture can gain sufficient stabilization from the hydrophobic effect, generally considered the dominant factor in protein stability^{1,2}. However, β -strands 8 and 10 of the three-stranded single-layer β -sheet make extensive interactions with the globular domains (Fig. 1), and thus these two strands

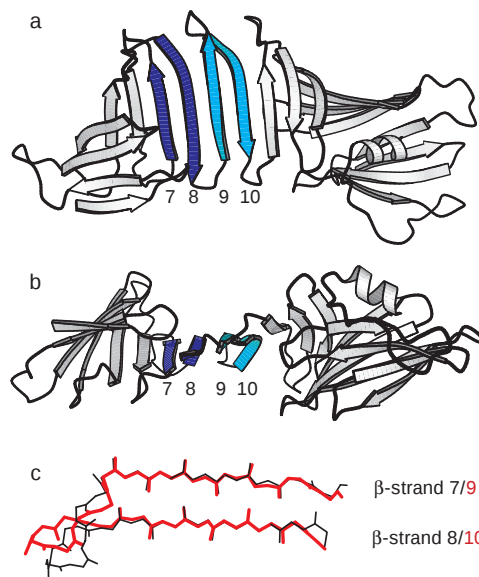


Figure 1 Structure of wild-type OspA. **a**, **b**, Drawings of the OspA structure⁷. The molecule in **b** is rotated along a horizontal axis relative to the molecule in **a**. The homologous β -hairpins (β -strands 7 and 8, and 9 and 10) in the single-layer β -sheet region are labelled and shown in blue and cyan, respectively. **c**, Superposition of the β -hairpin units. β -strands 7 and 8 (residues 95–118; black lines) and 9 and 10 (119–141; red lines) have been superimposed using C_{α} atoms in the β -strand regions. The root-mean-squared deviation for the backbone atoms in the β -strands is 0.51 Å.

could be considered as edge strands of the two globular domains. There are also a few interactions between the ninth, central, strand and the globular domains⁷. Thus, it could not be excluded that the single-layer β -sheet region is stably held only because it is a short segment tightly packed between two globular domains. Here, we examined whether interactions within the single-layer β -sheet region are sufficient to specify and sustain the β -sheet structure, by designing extended single-layer β -sheets and characterizing their structure and stability.

Two β -hairpins in and adjacent to the single-layer β -sheet region (β -strands 7 and 8, and 9 and 10, respectively) have homologous amino-acid sequences (Fig. 2), and their conformations are similar (Fig. 1c). This suggests that a β -hairpin unit similar to the existing ones could be inserted between β -strands 8 and 9, which would reproduce most of the original interstrand interactions (Fig. 2). Thus, we engineered OspA variants in which a β -hairpin segment (residues 119–141) was duplicated and triplicated (called OspA+1bh and OspA+2bh, respectively; Fig. 2).

We have completed the backbone NMR assignments of OspA+1bh (280 residues). Carbon-13 NMR chemical shifts of the C_α and C_β atoms are sensitive to the peptide backbone configuration, and thus they are accurate measures for identifying secondary structure elements¹². An analysis of ^{13}C chemical shifts of OspA+1bh revealed that the inserted segment (residues 119'–141') contains two β -strands as designed (Fig. 3b, c). The chemical-shift profiles of the two original β -hairpin units (β -strands 7 and 8, and 9 and 10) show only small changes upon the β -hairpin insertion in OspA+1bh (Fig. 3b, c). The chemical-shift profile of the inserted segment (strands 9' and 10') closely resembles that of strands 9 and 10. These results indicate that the conformation of the inserted segment may be similar to that of β -strands 9 and 10, and that the insertion did not significantly affect the conformations of the original β -strands (7, 8, 9 and 10) of the β -sheet. Furthermore, characteristic nuclear Overhauser effect (NOE) connectivities were detected in and near the insertion (Fig. 2), showing that the peptide backbone of the extended single-layer β -sheet is folded as designed. Residues in the globular domains of OspA+1bh show small or no chemical-shift deviations from those of wild-type OspA (data not shown), indicating small effects of the β -hairpin insertion on the globular domains. ^1H – ^{15}N NOE measurements provide information on the conformational dynamics of individual amide N–H bond vectors¹³. In OspA+1bh, residues that have reduced ^1H – ^{15}N NOE are localized in the terminal and loop regions, and the residues in the β -strands of the extended single-layer β -sheet have ^1H – ^{15}N NOEs close to the average (Fig. 3d), indicating that the region is not highly mobile on a picosecond–nanosecond timescale.

The ratio of the ^{15}N longitudinal and transverse relaxation times

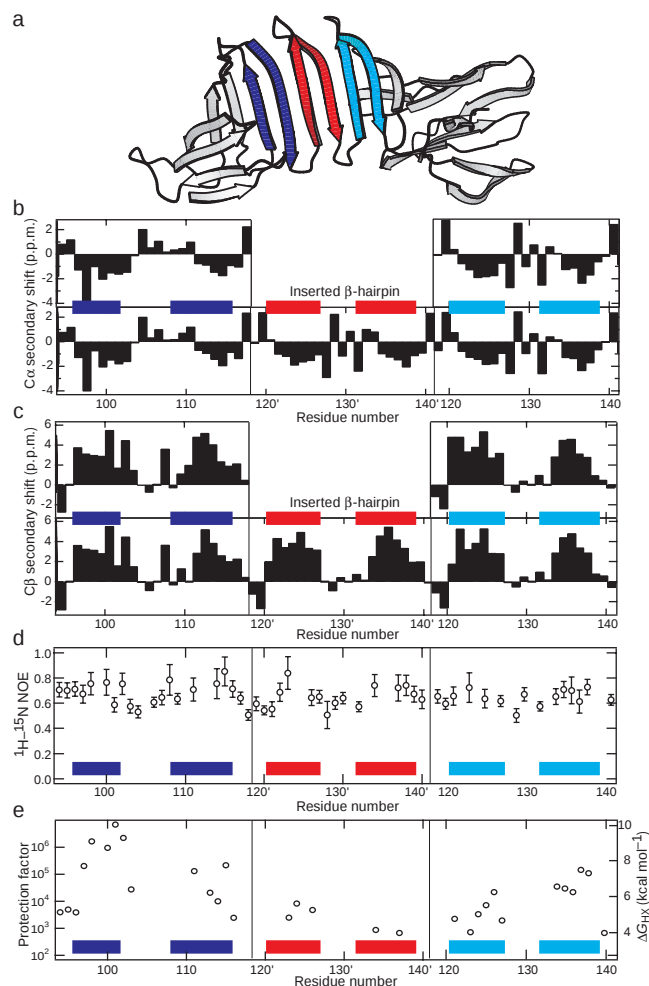


Figure 3 NMR characterization of the extended single-layer β -sheet in OspA+1bh. **a**, The model structure of OspA+1bh. **b**, **c**, Secondary shifts¹² of $^{13}\text{C}_\alpha$ (**b**) and $^{13}\text{C}_\beta$ (**c**) plotted against residue number. Consecutive residues with a $^{13}\text{C}_\alpha$ secondary shift larger than 0.7 p.p.m. with a $^{13}\text{C}_\beta$ secondary shift smaller than –0.7 p.p.m. define a β -strand¹². Upper rows show data for wild-type OspA, and the lower rows for OspA+1bh. Residue numbering is according to that in Fig. 2. Horizontal bars indicate the locations of predicted β -strands in OspA+1bh, and those of β -strands in the wild-type OspA crystal structure. **d**, ^1H – ^{15}N NOE intensities plotted against residue number. **e**, Amide HX protection factors¹⁸ ($P = k_{\text{intrinsic}}/k_{\text{observed}}$, where $k_{\text{intrinsic}}$ is the reference HX rate in the unstructured peptide; the scale is shown on the left vertical axis) plotted against residue number. The scale for the free energy of opening of a hydrogen-bonded amide proton ($\Delta G_{\text{HX}} = RT \ln(P)$) is shown on the right vertical axis.

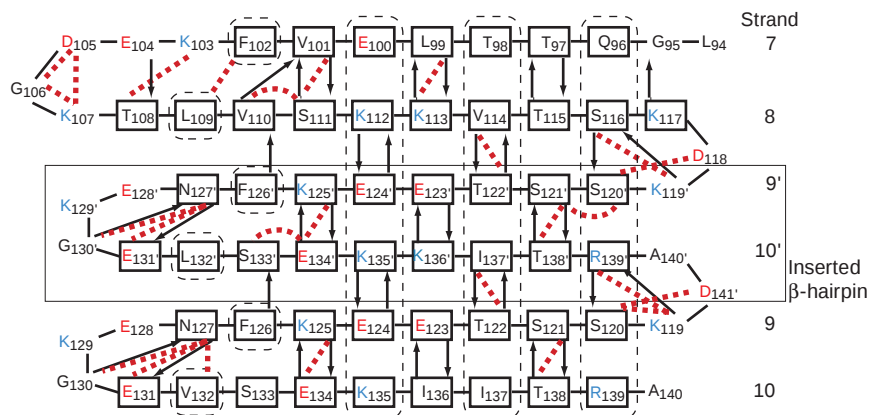


Figure 2 The β -sheet extension design in OspA+1bh. Residues 119'–141' denote those in the insertion. Two conservative mutations at residues 132' and 136' were introduced in the new β -hairpin to reduce the sequence identity between the β -strands 9'–10' and 9–10, and hence to increase the dispersion of NMR spectra. Residues in boxes are those

within β -strands; residues in dashed lines have side chains protruding toward the reader. Hydrogen bonds are shown as arrows, with the arrowhead pointing to the carbonyl group. Unambiguously observed $^1\text{H}^{\text{N}}$ – $^1\text{H}^{\text{N}}$ NOEs are shown as red dashed lines.

(T_1 and T_2 , respectively) is a good measure for the rate at which each N–H bond vector reorients because of global tumbling¹³, and it can be used to define long-range order in a protein that has a significant rotational diffusion anisotropy¹⁴. The ^{15}N T_1/T_2 ratios for 108 residues in OspA+1bh range from 14 to 50 (Fig. 4), indicating a large rotational diffusion anisotropy of the molecule. The data were best fitted with the equation describing the theoretical dependence of the T_1/T_2 ratio on the angle θ between the N–H bond vector and the long axis of the rotational diffusion tensor, when the diffusion axis is aligned parallel to the unique axis of the inertia tensor of the model structure (Fig. 4)^{10,14,15}. Data for residues from the amino- and carboxy-terminal globular domains have similar dependence on θ (Fig. 4), showing that the whole OspA+1bh molecule tumbles as a fairly rigid entity. Furthermore, many of the residues located in the β -strands of the single-layer β -sheet have large ^{15}N T_1/T_2 ratios that indicate slow tumbling. This is consistent with the model structure, in which the N–H bond vectors of these residues have small angles with the long axis of the molecule (Figs 2 and 3a).

Small-angle X-ray scattering (SAXS) has proved to be useful in defining the global conformation of a nonspherical molecule, such as OspA (ref. 9), and in distinguishing compact denatured states from a native protein¹⁶. The radius of gyration (R_g) for OspA+1bh was determined to be $28.4 \pm 0.2 \text{ \AA}$, in an excellent agreement with an estimated value of $28.5 \pm 0.1 \text{ \AA}$ from the model structure (Figs 3a and 5a). A radial Patterson ($P(r)$) curve, which is a real-space representation of the information in a scattering curve, shows the length distribution of atom-weighted interatomic vectors in the particle¹⁷. The $P(r)$ curve for OspA+1bh agrees well with that predicted from a model structure (Fig. 5a). In contrast, the $P(r)$ curve for a structural ensemble, in which the inserted segment is assumed to be a flexible linker, is distinct from the experimental data. Taken together, our NMR and SAXS data of OspA+1bh show that the extended single-layer β -sheet is formed as designed, that OspA+1bh is predominantly monomeric in solution, and that the insertion of the β -hairpin does not cause a significant bending of the β -sheet.

The R_g of OspA+2bh, the protein with two β -hairpin inserts, was determined to be $30.7 \pm 0.6 \text{ \AA}$, which is in good agreement with an estimation of $31.4 \pm 0.1 \text{ \AA}$ from our model. Its $P(r)$ curve also

agrees well with the prediction (Fig. 5b). We have not completed NMR characterization of OspA+2bh, because of technical difficulties associated with the larger molecular weight and the repetitive sequence in the single-layer β -sheet; however, the SAXS data indicate that the seven-stranded single-layer β -sheet may also fold into the designed structure. This is a critical test of the design, as the insertion of the second β -hairpin unit into the β -sheet of OspA+1bh should reproduce most of the interactions between the β -hairpin units that are already present in OspA+1bh (Fig. 2).

The degree of protection of amide protons from solvent exchange distinguishes a tightly packed protein from a loosely packed, molten-globule-like state^{3,18,19}. Amide hydrogen–deuterium exchange (HX) rates were determined for 117 amide protons of OspA+1bh at 45°C and pH 6.09. Each of the five β -strands of the extended single-layer β -sheet contains amide protons that exhibit substantial protection (Fig. 3e). Similar to the wild-type protein⁸, the most protected amides are located in the N-terminal globular domain. The free energies of ‘opening’ hydrogen bonds for the protected amides in strands 8, 9', 10', 9 and 10 range from 4.1 to $7.7 \text{ kcal mol}^{-1}$ at 45°C , indicating that the extended single-layer β -sheet is tightly packed. The relatively low protection factors for β -strands 9' and 10' indicate an increased mobility of the region. This might be due to suboptimum packing of the inserted segment caused by the two mutations at residues 132' and 136' (Fig. 2).

OspA undergoes a twostep unfolding reaction, in which the C-terminal folding unit, including β -strand 9 to the C terminus, unfolds first¹¹. Thus, the boundary between the two folding units coincides with that between the two homologous β -hairpins in the single-layer β -sheet (Fig. 2). Stability of the C-terminal unit can be specifically determined using fluorescence emission of the single tryptophan residue in this unit¹¹. A mutation in β -strand 9, F126A, decreases the unfolding free energy of the C-terminal folding unit by $2.0 \text{ kcal mol}^{-1}$ (Fig. 6), indicating that the stability of the OspA C-terminal folding unit is sensitive to perturbation in the single-layer β -sheet. Consequently, the stability measurement of the C-terminal folding unit can be used to probe the integrity of the single-layer β -sheet. The unfolding free energies of the C-terminal folding unit in the wild-type protein, OspA+1bh and OspA+2bh are $8.3 (\pm 0.2)$, $9.2 (\pm 0.2)$ and $8.8 (\pm 0.3) \text{ kcal mol}^{-1}$, respectively. The similar

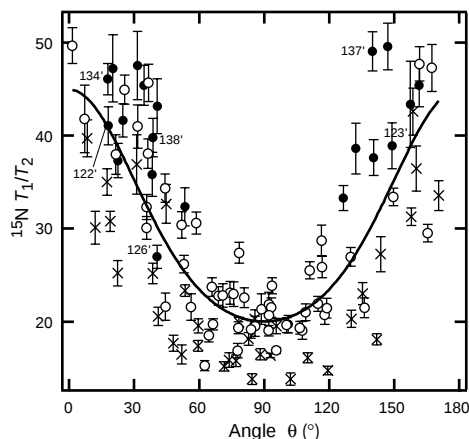


Figure 4 Nitrogen-15 T_1/T_2 ratios of the amide ^{15}N nuclei in OspA+1bh plotted against the angle θ between the N–H bond vectors and the estimated unique axis of the rotational diffusion tensor. Crosses indicate data for residues in the N-terminal globular domain, filled circles for those in the single-layer β -sheet, and open circles for those in the C-terminal globular domain. Data for residues in the inserted β -hairpin are labelled with residue numbers. The curve shows the best fit of the equation describing the θ dependence of the T_1/T_2 ratio^{14,15}.

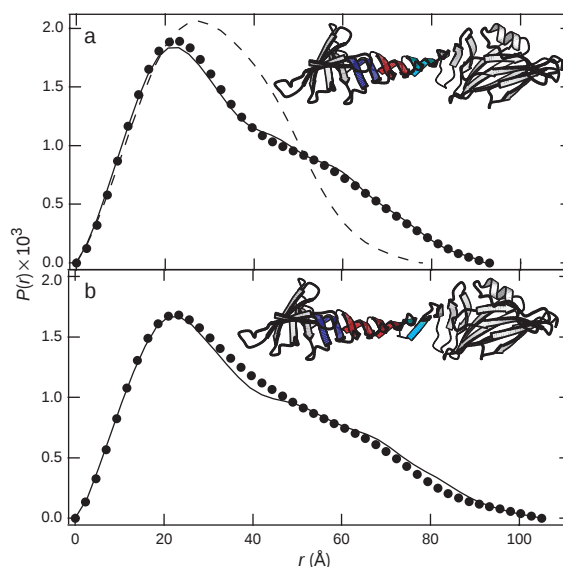


Figure 5 The $P(r)$ functions of OspA+1bh (a) and OspA+2bh (b) determined from SAXS data. Circles show experimental data, and the continuous lines are predictions from the models. In a, the broken line shows the $P(r)$ function for a structural model in which the inserted segment is assumed to be flexible. Models of the two proteins are also shown with the inserted β -hairpins in red.

stabilities for the three proteins show that the integrity of the single-layer β -sheet is maintained in the extended β -sheets. In contrast, an OspA+1bh variant containing three mutations in the inserted β -hairpin (K119'T/F126'L/K135'G) has a markedly reduced stability ($\Delta\Delta G^0 = 3.1 \text{ kcal mol}^{-1}$; Fig. 6), further confirming that the segment is an integral part of the protein. The dependence of the unfolding free energy on denaturant concentration (the m -value) was $2.44 (\pm 0.07)$, $2.88 (\pm 0.08)$ and $3.09 (\pm 0.11)$ for OspA, OspA+1bh and OspA+2bh, respectively. The m -value is known to correlate well with the change in solvent-accessible surface area upon unfolding²⁰; thus, the incremental increase in the m -value for OspA+1bh and OspA+2bh compared with that of the wild-type suggests that the size of the C-terminal folding unit increases as more β -hairpins are incorporated in the single-layer β -sheet, and that the insertion is part of the C-terminal folding unit.

Our successful design of large, stable single-layer β -sheets is, to our knowledge, the first example of the stable extension of a β -sheet in a natural protein. Further extension of the OspA β -sheet should be quite feasible, as no significant decrease in conformational stability occurred in OspA+1bh and OspA+2bh. Our results clearly indicate that interactions within the β -hairpin units, and between adjacent units, are sufficient for stability within the extended single-layer β -sheet structure. Thus, our results indicate that it may be possible to design a stable protein without the formation of a hydrophobic core. Peptides that form three-stranded β -sheets have been successfully designed²¹; also, the WW domain has been shown to fold as a three-stranded β -sheet^{22,23}. The high stabilities of the single-layer β -sheets in OspA variants are distinct from the marginal stabilities observed for these three-stranded β -sheets. This difference can be attributed to the fact that both edges of the OspA single-layer β -sheets are capped with globular domains that sequester additional surface areas of the single-layer β -sheet⁸. Thus, effective burial of the edge-strand surfaces may be important for the design of a small β -sheet motif with a high stability.

It is tempting to speculate that single-layer β -sheet structures may be involved in increases of β -sheet content that are associated with amyloid-like fibril formation and protein aggregation^{24,25}, because a single-layer β -sheet could be formed with a segment of a protein possessing a hydrophobic periodicity compatible with an amphipathic α -helix; such a β -sheet does not require the alternating hydrophobic/hydrophilic pattern of the amphipathic β -sheet. We have proposed that amino acids with long side chains play important roles in stabilizing the single-layer β -sheet structure by burying nonpolar surfaces⁸. It is intriguing that the 'polar zipper' sequences identified by Perutz *et al.*²⁶, such as the poly-glutamines associated with Huntington's disease, appear to satisfy the proposed requirements of stable single-layer β -sheets, suggesting potential involvement of single-layer β -sheet structures in human diseases.

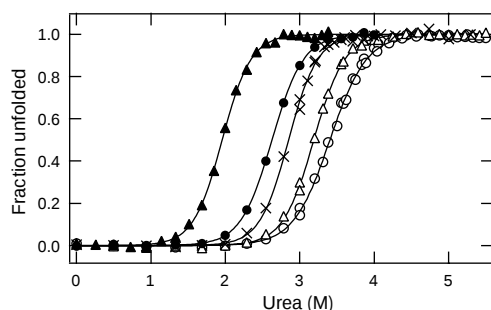


Figure 6 Urea-induced unfolding of wild-type OspA (open circles), OspA+1bh (open triangles), OspA+2bh (crosses), OspA mutant F126A (filled circles) and OspA+1bh mutant K119'T/F126'L/K135'G (closed triangles). Urea-induced unfolding reactions of the C-terminal folding unit in these proteins were followed using fluorescence emission from the single tryptophan residue in the C-terminal globular domain¹¹. The data were converted to fraction unfolded, according to the standard two-state model¹¹.

Methods

Sample preparation

The gene for OspA+1bh was constructed by inserting a 69-base pair (bp) DNA segment coding for the β -hairpin sequence (Fig. 2) in the wild-type OspA gene, using standard polymerase chain reaction (PCR) techniques. The second insertion was made by ligating a 69-bp DNA duplex at a unique restriction site (*Bse*RI) within the first insert. The F126A mutation was introduced using standard oligonucleotide-directed mutagenesis. Two unplanned mutations, K119'T and F126'L, were introduced by PCR errors during the construction of an OspA+1bh mutant, K135'G, which resulted in the triple mutations. The entire genes were sequenced to confirm that no other mutations had been introduced. The proteins were expressed and purified as described¹⁰.

NMR spectroscopy

The NMR sample contained 1.3 mM ^{13}C , ^{15}N -OspA+1bh in 10 mM sodium phosphate buffer, pH 6.0, with 50 mM sodium chloride in 95% H_2O and 5% D_2O . All NMR experiments were performed at 45 °C on a Varian Inova 600 spectrometer. Resonance assignments of OspA+1bh were achieved using standard triple resonance techniques, as described for wild-type OspA (ref. 10). ^1H - ^{15}N NOEs were identified in a 3D ^{15}N -edited NOESY spectrum²⁷. The chemical-shift data have been deposited in the BioMagRes Bank (accession number 4367). Carbon-13 secondary chemical shifts were calculated as the differences from respective random-coil shifts¹². Amide HX measurements were performed on a 1.0 mM ^{15}N -sample in sodium phosphate buffer (10 mM, pH* 6.09, the direct pH meter reading) containing sodium chloride (50 mM) at 45 °C, as described previously⁸. ^1H - ^{15}N NOE and ^{15}N T_1 and T_2 values were measured using published procedures^{10,28}. The N-H bond vectors were calculated from an energy-minimized structural model of OspA+1bh, as described¹⁰. The rotation diffusion property of this model structure can be well described with an axially symmetric diffusion model, in which the molecule is approximated as a prolate ellipsoid. The dependence of the T_1/T_2 ratio on the angle between the N-H vector and the unique axis of the diffusion tensor was analysed according to this model^{10,15}.

Small-angle X-ray scattering

The proteins were dissolved in sodium phosphate buffer (10 mM, pH 6.0) containing sodium chloride (50 mM). The protein concentrations were 5.2 and 5.15 mg ml⁻¹, for OspA+1bh and OspA+2bh, respectively. SAXS experiments were carried out at 30 °C as described⁹. Forty structures of OspA+1bh, in which the N- and C-terminal fragments were linked with a flexible linker corresponding to the inserted segment, were calculated using simulated annealing²⁹ with distance restraints that maintained the conformations of the terminal fragments.

Urea denaturation

Experiments were carried out in 10 mM sodium phosphate buffer (pH 6.0) with 50 mM sodium chloride at 30 °C, as previously¹¹.

Molecular graphics

We used Quanta (Molecular Simulations) for molecular modelling. The structures of OspA+1bh and OspA+2bh were modelled, in which the single-layer β -sheet was extended with an assumption that the inserted β -hairpin takes on a conformation similar to that of β -strands 9 and 10. Drawings were made using MOLSCRIPT³⁰.

Received 12 August; accepted 1 November 1999.

- Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.* **14**, 1–63 (1959).
- Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
- Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467 (1998).
- Beasley, J. R. & Hecht, M. H. Protein design: the choice of de novo sequences. *J. Biol. Chem.* **272**, 2031–2034 (1997).
- Bryson, J. W. *et al.* Protein design: a hierarchic approach. *Science* **270**, 935–941 (1995).
- Dahiyat, B. I. & Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).
- Li, H., Dunn, J. J., Luft, B. J. & Lawson, C. L. Crystal structure of Lyme disease antigen outer surface protein A complexed with an Fab. *Proc. Natl Acad. Sci. USA* **94**, 3584–3589 (1997).
- Pham, T.-N., Koide, A. & Koide, S. A stable single-layer β -sheet without a hydrophobic core. *Nature Struct. Biol.* **5**, 115–119 (1998).
- Bu, Z., Koide, S. & Engelman, D. M. A solution SAXS study of *Borrelia burgdorferi* OspA, a protein containing a single-layer β -sheet. *Protein Sci.* **7**, 2681–2683 (1998).
- Pham, T. N. & Koide, S. NMR studies of *Borrelia burgdorferi* OspA, a 28 kDa protein containing a single-layer β -sheet. *J. Biomol. NMR* **11**, 407–414 (1998).
- Koide, S. *et al.* Multi-step denaturation of *Borrelia burgdorferi* OspA, a protein containing a single-layer β -sheet. *Biochemistry* **38**, 4757–4767 (1999).
- Wishart, D. S. & Sykes, B. D. The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR* **4**, 171–180 (1994).
- Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28**, 8972–8979 (1989).
- Tjandra, N., Garrett, D. S., Gronenborn, A. M., Bax, A. & Clore, G. M. Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. *Nature Struct. Biol.* **4**, 443–449 (1997).

15. Lee, L. K., Rance, M., Chazin, W. J. & Palmer, A. G. Rotational diffusion anisotropy of proteins from simultaneous analysis of ^{15}N and $^{13}\text{C}\alpha$ nuclear spin relaxation. *J. Biomol. NMR* **9**, 287–298 (1997).
16. Kataoka, M., Kuwajima, K., Tokunaga, F. & Goto, Y. Structural characterization of the molten globule of alpha-lactalbumin by solution X-ray scattering. *Protein Sci.* **6**, 422–430 (1997).
17. Lattman, E. E. Small angle scattering studies of protein folding. *Curr. Opin. Struct. Biol.* **4**, 87–92 (1994).
18. Englander, S. W. & Kallenbach, N. R. Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q. Rev. Biophys.* **6**, 521–655 (1984).
19. Hughson, F. M., Wright, P. E. & Baldwin, R. L. Structural characterization of a partly folded apomyoglobin intermediate. *Science* **249**, 1544–1548 (1990).
20. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148 (1995).
21. Lacroix, E., Kortemme, T., de la Paz, M. L. & Serrano, L. The design of linear peptides that fold as monomeric beta-sheet structures. *Curr. Opin. Struct. Biol.* **9**, 487–493 (1999).
22. Macias, M. J. *et al.* Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **382**, 646–649 (1996).
23. Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel beta-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **8**, 841–853 (1999).
24. Fink, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.* **3**, R9–R23 (1998).
25. Kelly, J. W. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**, 101–106 (1998).
26. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl Acad. Sci. USA* **91**, 5355–5358 (1994).
27. Clore, G. M. & Gronenborn, A. M. Applications of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. *Prog. NMR Spectrosc.* **23**, 43–92 (1991).
28. Farrow, N. A. *et al.* Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry* **33**, 5984–6003 (1994).
29. Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
30. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950 (1991).

Acknowledgements

We thank B. M. Goldstein and J. J. Hayes for discussions; F. Delaglio, B. Johnson, L. E. Kay and A. G. Palmer III for NMR programs; and S. D. Kennedy for technical support. This work was supported in part by NIH grants to S.K. and D.M.E.

Correspondence and requests for materials should be addressed to S.K. (e-mail: Shohei_Koide@urmc.rochester.edu).

Going out on a LIMS

A round-up of lab information management systems and automation.

Tempo Pipet Tips

Zymark

www.zymark.com

Automation-Certified pipette tips

These now come in 100-ml volumes for use with Zymark workstations such as the Rapid-Plate 96/384. Tempo Pipet Tips have a specially designed attachment point for a reliable seal and accurate dispensing. The enhanced tip length reaches the bottom of deep-well plates, minimizing dead volume. They are available in 100- and 200-ml volumes, with or without filters and sterile or unsterile packaging.

Reader Service No. 100

Investigator Gel Processor

Genomic Solutions www.genomicsolutions.com

Eliminating manual procedures to enhance reproducibility of gel staining

This new processor is an automated system for staining and de-staining of multiple electrophoresis gels. The Investigator Gel Processor is a self-contained unit that can process multiple gels ranging in size from 8.25 to 25.25 cm. All reagents and liquid wastes are contained within the unit, which is mounted on wheels so that it can be easily rolled away and stored if not in use. Pre-programmed as well as customized protocols can be used in this device, including Coomassie blue, silver, SYPRO Ruby and other staining protocols.

Reader Service No. 101

AutoPrep-12

Hybaid

www.hybaid.com

Walk-away automation of DNA purification

With AutoPrep-12, Hybaid aim to automate one of the most widely performed procedures in molecular biology: plasmid DNA extraction and purification from crude bacterial cultures. The AutoPrep-12 extracts and purifies plasmid DNA from 12 cultures at a time. The resulting purified plasmid DNA is suitable for PCR, sequencing, and other typical downstream applications. Plasmid DNA extraction and purification is via alkaline lysis chemistry and patented filtration and purification steps, all in single-use columns.

Reader Service No. 102

KS Elispot 4.1

Zeiss

www.zeiss.com

Enhanced software for automatic evaluation of enzyme-linked immuno spot assays

Running under Windows 95 NT, KS Elispot 4.1 covers all processing steps, from image input and spot recognition to result presentation, printout, and data transfer. A redevelop-



Topical tips: Zymark's Tempo Pipet Tips.

oped teaching algorithm now performs the adaptation of parameter limits for spots by modifying the limits from the image itself, simply by clicking the mouse on the spots of interest. The resulting values may be saved and recalled for routine usage under a single file name. The evaluation is performed automatically using a Zeiss light microscope with a motorized stage and autofocus unit.

Reader Service No. 103

QuickStart

Beckman Coulter, Inc.

www.beckmancoulter.com

A preconfigured version of LabManager

This system offers all the major functions of the established LabManager C/S (client/server) LIMS but is pre-configured with the usual tables, fields, forms and reports needed in analytical laboratories. The LIMS can be further customized with standard LabManager C/S tools. Users can be productive in a fraction of the time it takes to build a fully configured LIMS. QuickStart runs via Windows.

Reader Service No. 104

MicroGrid

BioRobotics

www.biorobotics.com

A new microarraying robot

In designing this robot particular attention was paid to sample evaporation, carryover and run set-up processes. The system achieves a spotting resolution of 10 µm and submicrometre repeatability. Two versions are available: MicroGrid TAS for research groups who need wide flexibility in experiment design, and MicroGrid Pro for those requiring scaled-up production of a standardized microarray format. BioRobotics is taking orders for the first units, due to ship in April.

Reader Service No. 105

WinLIMS 4

QSI

www.qsius.com

LIMS for hand-held computers

Here's a laboratory information system that can operate in hand-held computers or organizers by virtue of direct operation with CE Windows. Work can be downloaded from the central system to allow sub-sets of work to be processed remotely and then automatically processed back into standard format. Operators can define samples and their associated methods, download that information from the LIMS and then move freely throughout a site entering data in the familiar Windows format of the central system. Barcodes can be entered and processed along with many other standard LIMS features according to the capabilities of the hand-held device.

Reader Service No. 106

Nautilus

LabSystems

www.labsystems.com

Entry-level LIMS

This 'ready-to-run' Nautilus LIMS simplifies set-up by offering LIMS preloaded on a Pentium III 500 MHz computer, the InfoMaker reporting package, Windows NT, Oracle RDBMS and peripherals. Consultancy, training and live installation are provided by LabSystems, including 90-day telephone support.

Reader Service No. 107

MiniTrak

Packard

www.packardinst.com

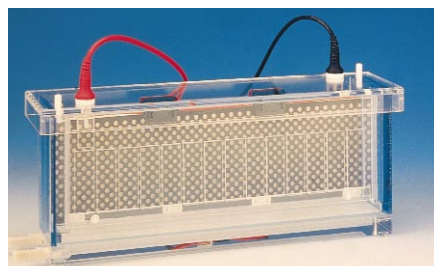
Automating microplate format applications

This four-module conveyor-based system is designed as an affordable approach to microplate processing. MiniTrak is a four-module conveyor system comprising two microplate stacker cassettes, a 96-tip pipette dispenser and an open module for future expansion. The system is compatible with 96-, 384- and other formats including deep-well



MiniTrak: automation on a microplate.

new on the market



Think tank: CBS's broad view on electrophoresis.

microplates. Unlike traditional automated pipettors, the open module provides flexibility and functionality via an optional 96- or 384-tip dispenser, wash module, vacuum filtration, lidding module or turbo aspirator/dryer. **Reader Service No. 108**

Electrophoresis and blotting

CBS Scientific www.cbsscientific.com

Triple-wide mini-vertical gel electrophoresis and blotting systems

Intended for mass sample screening, this system consists of two families of products: the Triple Wide Mini-Vertical Gel Electrophoresis Unit and the Triple Wide Mini-Blotter. The electrophoresis unit is available in either a single- or dual-sided configuration. The single-sided unit accommodates up to 102 samples and the dual-sided unit up to 204. This system is well suited for use with fluorescence-based assays since images can be captured directly through the optional borosilicate glass plates,

using a fluorimeter. The Triple Wide Mini-Blotter holds two 9.2 x 32.8-cm cassettes and can be used to transfer six standard minigels or two triple-wide gels simultaneously. **Reader Service No. 109**

Robotic liquid handling

Tecan www.tecan.com

A website focused on robotic liquid handling

Tecan invite users of their robotic liquid handling devices to join their User Group via the company's website. Information on the site is focused on Tecan's robotic liquid handling workstations and detection instruments, and their application in drug discovery, infectious disease screening and other life science and clinical diagnostic applications. **Reader Service No. 110**

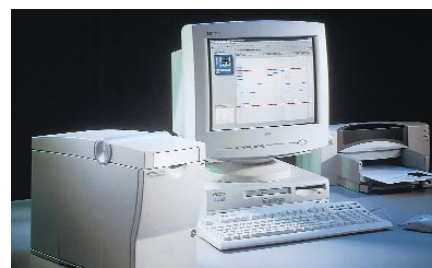
1500 Series Washing System

Heinicke www.hotpack.com

Microprocessor-controlled washing system

The Heinicke 1500 Series Washing System offers large washer capabilities in an upright package requiring only five square feet of floor space. Features include touch pad, a digital control package with preset wash cycles (heavy, medium and light), and three preset temperature settings, 55, 65 and 82 °C. Construction uses stainless steel throughout. The optional Heinicke 1550 Model comes with a high-watt, 230-volt water heater for faster heat-up times and throughput.

Reader Service No. 111



A chip off the old block: HP 2100 bioanalyser.

HP 2100 bioanalyser

Hewlett-Packard www.hp.com

Integrated analysis of nucleic acids

The HP 2100 make use of Caliper Technologies' LabChip technology. All sample preparation, fluid handling and biochemical analyses are carried out within a microchip containing microchannels, creating an interconnected network of fluid reservoirs. Used with a LabChip kit, the HP 2100 performs nucleic-acid analysis by integrating sample handling, separation, detection and digital data processing. The system is designed for use with PCR products, restriction enzyme digests or RNA preparations. Find out more about LabChip on www.hp.com/go/labonachip.

Reader Service No. 112

Links for LIMS

Contemporary Solutions www.csols.com

PC-based software for interfacing analytical instruments to LIMS

This software package is capable of linking any instrument to any LIMS in a bidirectional fashion. The latest version of the software includes a number of improvements designed to enhance operational performance in analytical practices such as duplicate analysis of samples. Particular attention has been paid to chromatography techniques, and for the use of balances in content and mass-uniformity measurements.

Reader Service No. 113

Autosampler syringes

SGE www.sge.com

Dual-gauge syringes for HP autosamplers

These tapered needle syringes of 0.5–100-ml capacity offer the strength of 23-gauge (0.63 mm) OD on the upper part of the needle, with a 26-gauge (0.47 mm) lower portion to allow the syringe to be used for split/splitless, direct, packed and on-column injection onto 0.53-mm ID capillary columns. Intended for use with Hewlett-Packard autosamplers, these syringes are available singly or in packs of six.

Reader Service No. 114

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card.

ADVERTISEMENTS

RT activity
the best assays
www.cavidi.com

READER ENQUIRY NO. 29

Peptides
818 residues
in multiple formats

Polyclonal Antibodies - 81,195

Please contact us about complete antibody and peptide services for your laboratory research and diagnostic applications.

Research Genetics
625, or Canada 800-333-4363
D.A. 0-800-88-3393
FAX 251-536-9116

READER ENQUIRY NO. 31

glycoSHOP.com

READER ENQUIRY NO. 32

FLOWERING OF A HYBRID

A discipline with its feet planted firmly in diverse sciences

p463

THE TERRITORY

The qualifications and characteristics necessary to succeed

p464

TISSUE ENGINEERING

Its uses have proliferated from a beginning in burns treatments

p466

BUILDING A CAREER

Company researchers welcome contact with students

p466

Bioengineering programmes rise to meet the challenge of a young science

When biology met engineering, they created a new field for scientists who don't want to be hedged into one discipline.

Scanning the c.v. of someone who is established in bioengineering (BE), you might get the impression of a highly educated individual who could not decide on a major field of study. "If you look at my c.v. I think this shows it well: materials science, mechanical engineering, mathematics," says Mauro Ferrari. After a maths degree in Italy, Ferrari was hired in 1989 by the University of California, Berkeley, as an assistant; he became a tenured professor after getting his graduate degree in mechanical engineering. He entered the faculty ranks with a split appointment between the departments of materials science and engineering and of civil and environmental engineering.

According to Ferrari, however, his work was always on micromechanics and its application to biomedicine. The problem back then, he says, was that the field was young and poorly defined. Berkeley did not have a biomedical engineering (BME) department. Instead, there were graduate groups — Ferrari joined those for BE, applied science and technology and biophysics.

Contrary to what his c.v. might suggest, Ferrari seems to know exactly where he is going and to enjoy the trail he is blazing. Today, he is getting closer to what he says he is supposed to be. In 1999, he moved to the Ohio State University (OSU) — still with a split appointment as a professor of internal



Ferrari: enjoys blazing a trail.

medicine and mechanical engineering, but now directing the OSU Center for Biomedical Engineering. He made the change to be nearer the interface of clinical work — UC Berkeley lacked clinical facilities. It's quite a leap, though, from physical scientist/engineer to clinical investigator. For example, says Ferrari, to get funded by the National Science Foundation, one must demonstrate 'proof of principle' for the technology. To receive funding from the US National Institutes of Health (NIH), one must show some clinical evidence. Currently, the mode of funding from the NIH is not conducive to triggering great advances in technology, he says.

Whither biomedicine?

According to Robert Nerem, director of the Georgia Institute of Technology's Parker H. Petit Institute for Bioengineering and Bioscience, the time has come to bring biology formally into engineering. First physics, then chemistry, were originally integrated into engineering. Biology is following.

Nerem makes an important distinction: biomedical engineering is simply the biggest

application of BE at the moment. "As we move into the twenty-first century, I think that there will be just as many applications outside the medical area where the integration of biology and engineering will be required," he says.

Douglas Lauffenburger, co-director of the Massachusetts Institute of Technology's (MIT's) division of Bioengineering and Environmental Health (BEH), agrees. He cites three reasons: biology is now mechanistic enough at a molecular level to be accessed by quantitative engineering analysis and design methodologies; new technologies built on biological components are becoming increasingly important to society; and complex biological systems require an infusion of systems-integrative engineering for continuing advances in fundamental understanding.

Elaborating on the broader meaning of BE, Lauffenburger sees a persistent confusion over the terminology, in that people use biomedical engineering and bioengineering interchangeably. Biomedical engineering is a field of application, he points out, whereas bioengineering is a new discipline: "We want to train students to apply it broadly to medicine, the environment, to materials and to defence." At the same time, he predicts that bioengineering will soon become the central discipline for major advances in biomedical engineering applications.

Passing the baton?

The Whitaker Foundation has long been a major supporter of biomedical engineering in the US. Peter Katona, president of biomedical engineering at the foundation, is hopeful that the NIH will show long-term commitment to bioengineering. While Whitaker has helped to build the educational infrastructure of biomedical engineering with some \$50 million annually, it now plans to shut its doors in 2006. At the moment, the foundation has about \$400 million, almost half of that already committed, says Katona. The rest will probably be committed in the next three years.

According to Katona, Whitaker research grants are typically intended for faculty members who are early in their careers at universities. The foundation also provides funds to universities to start programmes and departments in BME, which often include equipment, facilities and start-up costs for new staff.

With regard to the NIH, Katona hopes that some permanent arrangement will be found to continue what BECON is doing. This is extremely important, he says, because once Whitaker goes out of business he expects that BME will stand on its own. "NIH's recognition is a very big element of that," he says.

B.H.

Building research support

In 1997, NIH formed BECON, the bioengineering consortium. BECON has been the NIH-wide focal point for bioengineering, bringing visibility to their bioengineering needs and bridging the cultural gap between the BE community.

The NIH now seems eager to pursue BME. According to Dick Swaja, who works directly for Wendy Baldwin, NIH deputy director for extramural research and organizing chair for BECON, it is their overall responsibility "to ensure that the missions of the NIH are being met", while figuring out how BME will fit into NIH's traditional structure.

BECON concentrates its effort in four main

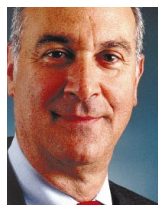
areas: nanotechnology and nano-science; tissue engineering (which includes biomaterials and, more specifically, biomimetics, biosensors, organ-culture systems and tissue regeneration); bioimaging; and bioinformatics and computer simulations, which is only now being used in medical teaching.

According to Swaja, BECON, or something like it, will continue to be the focus of activities at NIH for the near future. There is appropriations language that indicates that an Office of Bioengineering/Bioimaging, reporting to the director of NIH, will assume responsibility and direction of BECON.

In the mid-1990s, Nerem believed that it was enough to integrate BME into the various institutes at NIH. Over the past five years, however, "I've become convinced that unless you are an institute at NIH, you're a second-class citizen — the power really resides with the institute directors". For Swaja, the current environment is one in which there are opportunities for engineers and physical scientists to show what they are capable of doing and to demonstrate that they can work with the biomedical and imaging community to foster this transdisciplinary field.

Unconventionally NIH

BME feeds into a better understanding of cancer, as well as better treatment and diagnosis, says Carol Dahl, one of the National Cancer Institute's (NCI's) two primary representatives to BECON and director of the



Creative fusion of many minds

Working mainly with synthetic biomaterials such as lipids and polymers, as well as functional derivatives that involve carbohydrates and other molecules, James Baker (left) and colleagues at the University of Michigan, Ann Arbor, on the Unconventional Innovations Program, are creating devices that will engineer cells on a nanoscale.

Polymers, for example, can make 'smart delivery' systems for drugs and genes. They target a malignant cell, release the drug or gene and report any effects.

The participation of people from liberal arts and sciences, chemistry, physics, engineering and toxicology is crucial, says Baker, who is head of allergy and immunology in the medical school as well as leading the university's centre for biologic nanotechnology. "All of us have come to this, if I can put it politely, 'ad hoc', training ourselves," he says. "There is a lot of self-selection in this area because people who don't want to interact won't show up." He explains that his team can define something experimentally and then say to the synthesis people: "We see this structure-function relationship, how can we magnify that?" The toxicologist might then say the compound has a bad history *in vivo*, vetoing that line of thinking. "It's remarkable how you can overcome problems or prevent yourselves from going down blind alleys, if you have everyone together."

Baker and his colleagues would like to see a larger institute based on their current centre. "One of the really important things is that you get people who are cross-trained and understand the engineering side, the materials science side and the analytical side, as well as the biology and the toxicology," he says. **B.H.**

Office of Technology and Industrial Relations at the NCI. To capitalize on this opportunity, the NCI has been an active participant in BECON and its initiatives, as well as developing NCI-specific programmes to nurture technology development and BE.

These have been created to smooth the problems technology and BME have encountered in gaining support from what is primarily a hypothesis-driven research focus at the NIH, says Dahl. The Phased

Innovation Award, developed by the NCI, allows people to move rapidly from a feasibility stage to a development stage (see NCI in the 'Institutions' panel on page 466). There is also the Unconventional Innovations Program (UIP), an NCI programme that invests in long-term, high-risk and high-impact areas.

The NCI takes a more active scientific role in the UIP, which announced its first five awards last autumn (see 'Institutions' panel for new announcements). Investigators

Exploring the territory in tissue engineering

A disagreeable side effect of longer lifespans is the failure of one part of the body — the knees, for example — before the body as a whole is ready to surrender. For decades, bovine collagen and other materials have been used for repairs, but it is now possible to use human material to restore damaged or worn-out tissue.

Tissue engineering, which occupies the fertile area between materials science and biology, has its roots in cell biology, immunology, chemistry and bioengineering, but it's not uncommon to see apparently unrelated fields thrown together. Glenn Booma, a director of 'outcomes research' at Genzyme Tissue Repair, in Cambridge, Massachusetts, cites a laser-optics physicist joining a team of physiologists working on a process for solidifying an injected substance by photoactivation.

A working knowledge of more than one branch of science could be very helpful. A postgraduate education is necessary for most research jobs, although a bachelor's degree may suffice in technical support roles. Equally important is an ability to work and communicate with people from other fields, and to keep a global perspective

even when concentrating on the details of a project. In a company environment the ability to shift gears rapidly is also a must. If a project turns out not to work, or if a competitor files a patent, says Booma, "then the company is going to want to rapidly reassign you to something new".

Your fate is out of your hands

Nancy Parenteau, chief scientific officer at Organogenesis, a company making wound-repair products in Canton, Massachusetts, says that the cross-fertilization between disciplines is part of the fun. "You should be able to pick up some of the other specialty's language, appreciate their issues and how they go about doing things." The hardest thing for scientists to come to grips with, she says, is that a project does not depend solely on any individual's work, even if it's an outstanding achievement. "In a way, your fate is no longer in your own hands. That can be difficult to become comfortable with."

Non-scientific factors can impinge on careers in a small company. Booma says that none who saw the seminal paper on cartilage defects in rabbits in 1984 would have dreamed it would take until 1997 to get

approval from the Food and Drug Administration and commercialize their Carticel product. Now, however, investors do know, so even small companies with impressive technology platforms and good intellectual property may be bought up, as his company was in 1994 when, as Biosurface Technology, it joined Genzyme. "Genzyme and Organogenesis have shown that tissue-engineering products are effective and marketable, but now we have to show that they can be profitable," he says.

Those contemplating a career in tissue engineering should also be politically aware. Positive public attitudes translate into public support, whereas a negative impression can mean restrictive legislation, loss of funding and, in extreme cases, physical attacks on labs and personnel. Scientists should engage in discussion and clarify issues. Parenteau deplores the speculation that occasionally appears in the media as being harmful to the field and misleading and frightening to the public. Even preliminary or limited results can be misinterpreted, raising false expectations for many and confusing the few for whom a breakthrough will truly bring help. "The

propose the work and the NCI then negotiates with them and engages them under contract for the next few years. The five most recent contracts amount to nearly \$11.3 million over three years.

James Baker, principal investigator for a group receiving \$4,427,711 at the University of Michigan at Ann Arbor, will use the grant to develop nano-scale devices for detecting and treating cancer. Although there was a lot of input from the NCI, says Baker, "I have to give them their due. They are setting their priorities, but are allowing us to direct towards those priorities in a way that makes the most sense given the specific technology."

One common aspect among the BE and BME groups that *Nature* has spoken to is that they have organized their departments and divisions in ways that are administratively more challenging in order to gain an intellectually richer environment.

At MIT's division of bioengineering and environmental health, most faculty members hold either 'dual' or 'joint' appointments. "We believe that joining the science of biology with engineering is so important that you do not want to do it in an isolated structure that departments often are," says Lauffenburger. Dual appointments are shared equally with the other department in terms of teaching and service duties, as well as hiring, promotion, tenure and salary review. Joint appointments are positions in which these areas are

governed by another department. The departments sharing BEH faculty include chemical, mechanical and electrical engineering, as well as computer science, materials science, biology and chemistry. "Our strategic plan calls for adding another 10 to 12 faculty over the coming five to seven years, from a combination of internal transfers and external hiring," says Lauffenburger.

At the graduate level, BEH offers PhD degrees in bioengineering and toxicology. At the undergraduate level, there are minor bachelor degrees in both biomedical engineering and environmental health, and a mechanical engineering degree is in the planning stages. The goal of the graduate-level programmes is to define these disciplines, whereas the undergraduate programmes will connect traditional disciplines with problems and approaches in biotechnology and human health.

At the Georgia Institute of Technology (GT), a joint biomedical engineering department is being put together with the medical school at nearby Emory University. According to Nerem, the department will have space on both campuses, with two-thirds of faculty appointed through GT and one-third through Emory. Once appointed, they will act as a single department. A new PhD degree is going through the approval process, which will be a BME degree offered jointly by GT and Emory.

Over the next seven or eight years, Nerem

estimates that GT will be hiring some 15 faculty. There are already seven faculty with primary appointments in the department, he says. They did not want simply to move the existing faculty into the new department. "I think it's important that mechanical engineering at GT is still strong in biomedical research," he says. A few of the faculty have moved into this new department, but most will be new faculty.

Generally speaking, Nerem thinks that "some of the best engineering students have an interest in biology and are excited about bioengineering". Getting biology students interested in engineering is not quite so easy, he says. The problem, he explains, is that many people who do biology as an undergraduate shy away from quantitative courses. Too often biology majors do not study calculus or ordinary differential equations.

According to Lauffenburger, bioengineering is in the very early stages of development in terms of both its definition and the identification of career opportunities for its students. "I think that it's a crucial challenge for us to help industry understand what we're trying to accomplish with this education, and how people like this will be useful to hire alongside the chemical, mechanical and electrical engineers they have traditionally hired," he says.

Brendan Horton

Brendan Horton is a golfer and freelance science writer.

e-mail: b.horton@naturedc.com

companies themselves, which have to go out and raise money in order to survive, have a need to play up the bright side and not dwell unduly on the limitations," Booma adds.

New university training programmes and faculty positions and public funding initiatives are good news for hiring trends. In the United States, the National Institute of Standards and Technology's Advanced Technology Program has funded product development in tissue engineering for the past two years with grants of \$2–5 million. Last year the NIH set up a tissue-engineering working group, and may set up an institute. Says Gail Naughton, president of Advanced Tissue Sciences in La Jolla, California, "Tissue engineering used not to fit into any single category, but now the agencies are actually out there soliciting grant proposals. This is making a big difference." She foresees career opportunities at various levels.

With growing institutional support, the outlook for the next 10 to 15 years is rosy. Improvements in technique and understanding will speed up progress. For example, Naughton thinks that the question of embryonic stem cells will be obviated by improvements in separations and culturing.



Parenteau: hybrid vigour adds interest.

"Our growing understanding of how to keep cells functioning outside the body will allow us to manipulate them so that they can become almost any cell type," she predicts, citing recent publications (for example, see Jackson, K. A. *et al. Proc. Natl Acad. Sci. USA* 96, 14482; 1999) showing that human cells collected from bone marrow can be transformed into fat, bone or cartilage cells.

Parenteau worries about speculative media coverage, but agrees that what's true

now for skin products will probably become reality for vascular, neurological, bone, tendon and corneal tissue. The immunological barrier is a major challenge, but inroads are being made. For example, the Boston company Diacrin has worked out how to mask animal cells so that they can be transplanted into humans; a product made from porcine neural cells to treat Parkinson's disease is in phase II trials. A decade ago neural regeneration at trauma sites was thought impossible. Today it is known that, if scarring is kept at bay, neural tissue has a great capacity to regenerate.

The role of genetics will also increase as tissue engineering is combined with gene therapy to enhance the growth of implanted tissue. The discoveries of genomics will affect the field, such as the as-yet unknown signals that differentiate between normal growth and healing. "Especially in a young field the opportunities are broad and abundant. It is exciting to participate in a field that is being born. Few of us get to do that," says Parenteau.

Potter Wickware

Potter Wickware is a science writer in Mill Valley, California.

e-mail: wickware@worldnet.att.net

Progress from a fragile start

Tissue engineering was born at MIT in the late 1970s with Howard Green's work on epithelial sheets and keratinocyte grafts. At the same time, Ioannis Yannas was in another department working on collagen scaffolds for dermal repair. Robert Langer was looking at synthetic polymers, originally for drug delivery, then later, with Eugene Bell, redirected towards cell lattices. Yannas was one of the first to try an implant that would direct tissue formation. He worked with a burns doctor, John Burke, at the Shriners Hospital in Boston and, the technique was initially called the Burke-Yannas method.

Before the mid-1990s tissue engineering was confined to autologous-cell-based skin products for treating burns. The cells were taken from a plug of the patient's biopsy material and used to grow keratinocytes in sheets. The material took well, the cosmetic result for the patient was good, and it was a major advance over cadaver or pig skin, used as temporary covering for years. It then began to be used for chronic wounds, particularly leg and foot ulcers. There were problems, though: the fragile, single-layer sheets were prone to blistering and tearing, and culturing took up to three weeks.

Since then, progress in the multiple disciplines that make up tissue engineering have led to more substantial composite skin materials (see 'Companies' panel above). They are grown continuously in a factory-like setting, are available at short notice, have a reasonably long shelf-life and are relatively easy to handle. Advanced Tissue Sciences' skin product, TransCyte, consists of allogeneic fibroblasts

Companies

- | | | |
|--|---|---|
| ● Advanced Tissue Sciences
www.advancedtissue.com | ● Genzyme Tissue Repair
www.genzyme.com/company/lines/tissue/welcome.htm | ● Organogenesis
www.organogenesis.com |
| ● Chiron
www.chiron.com | ● Layton Bioscience
www.laytonbio.com | ● Scios
www.sciosinc.com |
| ● Clonetics
www.clonetics.com | ● LifeCell
www.lifecell.com | ● SyStemix
www.stem.com |
| ● Diacrin
www.diacrin.com | ● One Cell Systems
www.onecell.com | ● TissUse
www.tissuse.com |
| ● FibroGen
www.fibrogen.com | | ● VITA34
www.VITA34.de |

— cells not from the patient's own body — seeded into a polyglycolic-acid mesh. This material does not engraft into wounds, but the fibroblasts cover the wound effectively and emit wound-healing signals. Pain is reduced, as is the need for dressing changes and nursing care. It has been approved for partial thickness burns and is in trials now for venous ulcers and foot ulcers.

Another product, Organogenesis's Apligraf, consists of a type I collagen matrix seeded with allogeneic fibroblasts and covered with a layer of cultured allogeneic keratinocytes, simulating dermal/epidermal tissue. It was approved for treating venous ulcers a year and a half ago, and is in review for diabetic foot ulcers.

At least as important as material strength and durability are the serological properties of the tissue-engineered materials. Allogeneic cells, derived from neonatal prepuces, grow rapidly in culture, stimulate wound-healing and, perhaps most importantly, are immunologically neutral. Moreover, the keratinocytes and fibroblasts, being derived from immature tissue, and having gone through several passages in cell culture, do not bear the HLA class I and II surface proteins that enable the immune system to carry

out 'self/non-self' recognition surveillance and initiate the disastrous graft-rejection response.

Wound-healing signals emitted by engineered tissue are believed to be a combination of cytokines and specific growth factors. Bringing these complicated interactions into focus is an active area of research in laboratories around the world.

Potter Wickware

Insiders advise on first steps to a career

Jim Polarek directs collagen development at FibroGen, a Southern San Francisco company that provides cultured human collagen for matrix to companies and research groups. He advises anyone interested in tissue engineering to consider an internship, either for an undergraduate summer job or as a postgraduate. A biologically oriented internship can be especially valuable for students in engineering disciplines, he adds. Gail Naughton, of Advanced Tissue Sciences,



Naughton: interns are welcome.

points to her company's programme, which supports up to 24 graduate and undergraduate students.

Other useful contacts are the Tissue Engineering Society, the Pittsburgh Tissue Engineering Initiative and company scientific advisory boards. Glenn Booma of Genzyme

says, "We're more than happy to have people contact us and come in and take a tour; it's one of the ways we find good people."

Booma adds that for management jobs a good combination is dual master's degrees: one a BS/MS from a five-year programme followed, after a few years' work experience, by an MBA. "The niche specialist in today's job market is challenged, but a dual degree shows you can have a broad understanding," he comments.

Potter Wickware

Institutions

- | | | |
|--|--|--|
| ● BECON, NIH
grants.nih.gov/grants/becon/becon.htm | Network
www.bmenet.org/BMEnet/Whitaker | Physical Science
www.nih.gov/od/ors/beps |
| ● Biomedical Engineering Graduate Schools
www.gradschools.com/listings/menus/biomed_eng_menu.html | ● European Tissue Repair Society
www.leahcim.demon.co.uk/etrs.htm | ● NSF, Bioengineering and Environmental Systems
www.eng.nsf.gov/bes |
| ● Biomedical Engineering Home Pages
www.whitaker.org/academic/biomed-homes.html | ● Gesellschaft für Biotechnologische Forschung, mbH
www.gbfbraunschweig.de | ● National Institute of Standards & Technology
www.atp.nist.gov/atp/focus/tissue.htm |
| ● Biomedical Engineering Program Industry Survey
www.lance.colostate.edu/depts/bme/industry/industsurv.html | ● The International Federation for Medical and Biological Engineering
www.ifmbe.org | ● Pittsburgh Tissue Engineering Initiative
www.pittsburgh-tissue.net |
| ● Biomedical Engineering | ● University of Leipzig
http://www.uni-leipzig.de/~ikit/index.htm | ● Tissue Engineering Society
www.ummed.edu/dept/tes/ |
| | ● NIH, Bioengineering and | ● National Cancer Institute Unconventional Innovations
amb.nci.nih.gov/RFP/07013/07013.htm |